High Resolution Environmental Modelling Application Using a Swarm of Sensor Nodes

Ferry Susanto

College of Engineering and Science

Victoria University

This dissertation is submitted for the degree of

Doctor of Philosophy

Engineering and Science

March 2017

I dedicate this thesis to my loving parents.

Declaration

I, Ferry Susanto, declare that the PhD thesis entitled "High Resolution Environmental Modelling Application Using a Swarm of Sensor Nodes" is no more than 100,000 words in length including quotes and exclusive of tables, figures, appendices, bibliography, references and footnotes. This thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma. Except where otherwise indicated, this thesis is my own work.

Ferry Susanto
March 2017

List of Publications

The following are publications of work undertaken as part of this thesis:

1.

Design of Environmental Sensor Networks
Using Evolutionary Algorithms

Ferry Susanto, Setia Budi, Paulo de Souza Jr., Ulrich Engelke, and Jing He

Title : **Design of Environmental Sensor Networks Using Evolutionary Algorithms**

Authors : **Ferry Susanto**; Setia Budi; Paulo de Souza; Ulrich Engelke; Jing He

Journal : IEEE Geoscience and Remote Sensing Letters

IF 2015 : 2.228

Date : 26 February 2016

doi : 10.1109/LGRS.2016.2525980

2.

SENSORS

Spatiotemporal Interpolation for
Environmental Modelling

Ferry Susanto, Paulo de Souza Jr., and Jing He

Title : **Spatiotemporal Interpolation for Environmental Modelling**

Authors : **Ferry Susanto**; Paulo de Souza; Jing He

Journal : MDPI Sensors

IF 2015 : 2.033

Date : 6 August 2016

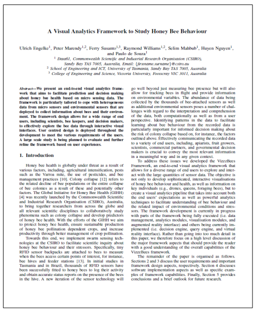
doi : 10.3390/s16081245

I also co-authored the following works during the progress:

1.

Title	:	**Visual Assessment of Spatial Data Interpolation**
Authors	:	Ulrich Engelke; **Ferry Susanto**; Paulo A. de Souza Junior; Peter Marendy
Conf.	:	Big Data Visual Analytics (BDVA), 2015
Date	:	22-25 September 2015
doi	:	10.1109/BDVA.2015.7314305

2.



Title	:	**A Visual Analytics Framework to Study Honey Bee Behaviour**
Authors	:	Engelke U, Marendy P, **Susanto F**, Williams R, Mahbub S, Nguyen H, and de Souza P
Conf.	:	Proc. of IEEE International Conference on Data Science Systems (DSS), Sydney
Date	:	December 2016
doi	:	10.1109/HPCC-SmartCity-DSS.2016.0214

3.

Title	:	**In Search for a Robust Design of Environmental Sensor Networks**
Authors	:	Setia Budi, **Ferry Susanto**, Paulo de Souza, Greg Timms, Vishv Malhotra and Paul Turner
Status	:	Accepted - Environmental Technology
Date	:	18 Mar 2017
doi	:	10.1080/09593330.2017.1310303

List of work-in-progress publications:

1.	Title	:	**Inferring *Apis mellifera* Behaviour from Population Activity**
	Authors	:	**Ferry Susanto**, Thomas Gillard, Paulo de Souza, Benita Vincent, Setia Budi, Auro Almeida, Gustavo Pessin, Helder Arruda, Raymond N. Williams, Ulrich Engelke, Peter Marendy, Pascal Hirsch
	Status	:	Manuscript completed - Elsevier, Ecological Modelling
2.	Title	:	**Data-driven Field Simulation and Environmental Modelling for Swarm Sensing Project**
	Authors	:	**Ferry Susanto**, Paulo de Souza Jr., Raymond Williams, Thomas Gillard, Setia Budi, Peter Marendy
	Status	:	Manuscript completed - IEEE Transactions on Geoscience and Remote Sensing
3.	Title	:	**Agent-based Modelling of Honey Bee Forager Flight Behaviour for Swarm Sensing Applications**
	Authors	:	Paulo de Souza, Raymond Williams, Stephen Quarrell, Setia Budi, **Ferry Susanto**, Benita Vincent, Geoff Allen, Auro Almeida, Dale Worledge, Leandro Disiuta, Pascal Hirsch, Gustavo Pessin, Helder Arruda, Peter Marendy, Leon dos Santos, Tom Gillard, Andojo Ongkodjojo Ong
	Status	:	Under review - Environmental Modelling and Software
4.	Title	:	**Design of Environmental Sensor Networks with Support of Mobile Platform**
	Authors	:	Setia Budi, Paulo de Souza, Greg Timms, **Ferry Susanto**, Vishv Malhotra and Paul Turner
	Status	:	Manuscript preparation.

5.	Title	:	**Low-Cost Electronic Tagging System for Bee Monitoring**
	Authors	:	Paulo de Souza, Benita Vincent, Stephen Quarrell, Gustavo Pessin, Setia Budi, **Ferry Susanto**, Geoff Allen, Peter Marendy, Auro Almeida, Dale Worledge, Pascal Hirsch, Leandro Disiuta, Ulrich Engelke, Huyen Nguyen, Raymond Williams
	Status	:	Manuscript preparation.

Acknowledgements

This project would not be possible without the support of many people. First of all, this thesis is dedicated to my parents, who have unconditionally supported and encouraged me. I would like to thank a number of institutions that have financially supported my study: (i) Victoria University for waiving my PhD tuition fees; (ii) Vale Institute of Technology for the award of a postgraduate scholarship; and (iii) CSIRO's the Office of the Chief Executive program for a top-up scholarship. This wonderful support has allowed me to fully focus entirely on my PhD work.

I express my deepest gratitude to my supervisors: Prof. Jing He (Victoria University, Melbourne), Prof. Paulo de Souza (CSIRO), and Guang Yan Huang (Deakin University), for the continuous support and guidance that they have offered throughout the PhD progress, as well as for their patience, enthusiasm, encouragement, and knowledge. Their feedback on the research work and the thesis writing have contributed greatly to the success of this work.

I would also like to thank CSIRO for providing four different data sources to allow the development and experimental simulation possible, they are: (a) a 'modelled' Environmental data of the South Esk region, in Tasmania; (b) a 'benchmark' data set that collects environmental observations from a number of weather stations at South Esk region; (iii) an agent-based bee flight simulator developed by the Swarm Sensing team; and finally (iv) real bee experimental data using RFID-based systems conducted at Geeveston, Tasmania.

Furthermore, I want to express my appreciation to CSIRO's Swarm Sensing team members and colleagues who have helped me in a number of ways such as research development, data set preparation, understanding bee biology, undertaking administrative work throughout my PhD. These colleagues include Setia Budi, Ray Williams, Tom Gillard, Benita Vincent, Stephen Quarrell, Ulrich Engelke, Peter Marendy, Pascal Hirsch, and Selim Mahbub. Their help has accelerated my learning process, and also greatly improved the quality and presentation of this dissertation.

Abstract

The advancement of sensing technology has successfully reduced the physical size of a sensor node and stimulated the application of swarm sensing (millimetre scale sensors). Such a system has been envisioned to provide novel applications. For example, CSIRO has commenced the application of swarm sensing technology to track insect that aims to understand how the environmental situation influences bee behaviour.

While the micro-sensor is still under development, it is crucial to have a baseline data set for initial data analysis purposes so that reasoning with the rich data is possible once the hardware has been developed and deployed. This work will propose a field simulation to address this issue. A hybrid environmental sensor network will be deployed, for the purpose of making highly dense observations, that consists of: (i) fixed sensor nodes, acting as weather stations, that collect data in a regularly-spaced time interval; and (ii) mobile nodes that sense the environmental parameters while insects move within the region with extremely high frequency – i.e. seconds.

The proposed spatio-temporal interpolation algorithm in this dissertation (i.e. for environmental modelling) has achieved a computational efficiency factor from highly-dense sample data with an acceptable statistical error. The method also reconstructs the environmental situation in reality – e.g., produce a smooth surface in space and over the time.

Combination of a successfully developed field simulation and the interpolation algorithm has opened up a wide range of applications. For instance, researchers could infer bee behavioural dynamics based on the environmental changes that they are experiencing. Such activities could assist entomologists to deepen their understanding of bee behaviour, with a view to advance our knowledge of the decline in bee populations worldwide.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Background

Wireless sensor network is a set of interconnected sensor nodes deployed with the purpose of observing environmental conditions such as temperature, relative humidity, solar radiation. Such a system has become the key technology for the future of environmental monitoring, allowing us to observe environmental variables at difficult and hostile locations such as mountainous and deep marine region. This technology offers a significant contribution to our society, for instance, early warning system, weather forecast, and precision agriculture. However, the deployment of those fixed sensor nodes (e.g., weather stations) offer a low spatial coverage. Also, it requires an appropriate deployment strategy (also known as *spatial sampling* method) to obtain a cost-effective and a fit-for-purpose network (i.e., a required level of network representativeness).

A mobile sensor node, on the other hand, is more versatile than fixed sensor node. It makes observations while it moves through the area under study (e.g., aerial and autonomous underwater vehicle). It suffers from the limitation of having a small temporal coverage at discrete locations. With the advancements in the sensor technology, the cost and size of an individual sensor node have been decreased. For instance, the Smart Dust project [1] was envisioned to provide thousands of millimetre-scaled sensing with high spatial and temporal coverage (also known as *swarm sensing*). Such a network provides much more discrete measurements of the environment in a less intrusive way. Example applications using such a technology could be found in the military (e.g., unmanned aerial vehicles) and animal (or even insect) tracking.

Irregularly distributed data is obtained from the sensor network which requires estimation of the missing data at locations without measurements. Spatial or spatiotemporal interpolation method can be utilised (environmental modelling) to allow environmental managers to obtain

a continuous surface of the region under study, in which able to support a more accurate interpretation and decision making. Nonetheless, choosing a suitable interpolation method is a non-trivial task, and an in-depth investigation must be carried out for the applications in different purposes. This is because the performance of any interpolator can vary based on different data sets and requirements (e.g., accuracy and computation efficiency).

1.2 Motivation

CSIRO (Commonwealth Scientific Industrial Research Organisation) has commenced the development of the swarm sensing technology that attaches micro-sensor to insects (bee in this case). Such a configuration allow the mobile sensor nodes to make observations as the bees moving throughout the Region of Interest (RoI). As a result, highly dense data sample will be obtained, this has stimulated the demand for a computational method to process and visualise a large amount of data and to make sense of them. CSIRO's Swarm Sensing Project (SSP) aims to better understand bees' response to different stressors such as environment (e.g., strong wind, heavy rain, extreme weather, pollution) and the surroundings they are experiencing in (e.g., pesticide exposure, varroa mites, human intervention).

Fig. 1.1 An example of a bee with a Radio Frequency Identification (RFID) micro-sensor (2.5mm × 2.5mm × 0.4mm, 5.4mg) mounted on its thorax. Credit: CSIRO.

The initial bee experiment conducted by CSIRO involved installing a Radio-Frequency Identification (RFID) reader at the hive entrance and glueing a RFID tag is being glued on each bee's thorax (Figure 1.1), so that the presence of a bee passing through the entrance is stored in a database. The experimental set-up, resulting from current limitations in technology, neither allow us to track individual bee flights paths nor provide us information whether an insect is outside or inside the hive. However, such a data set has offered us an opportunity to

utilise computational techniques to infer bee behaviour at colony level so that we can answer the following question: what is the probability of an individual bee in the colony to forage at a particular time of day (e.g. 6am, 1pm and 4pm)?

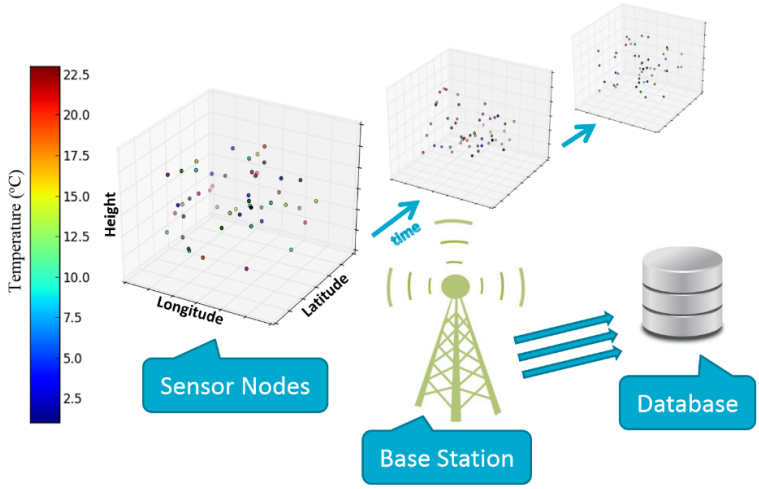


Fig. 1.2 An illustration of sensor networks to be developed within the SSP: (a) *Sensor nodes*. An individual sensor that records environmental data, i.e., bees in this case; (b) *Base station*. Infrastructure acting as an "agent" that receives values from individual sensor nodes and sends them to the database to be stored; (c) *Database*. A medium that records the raw data collected, which always comes with a periodic backup system for security purposes.

Figure 1.2 illustrates the sensor network design to be developed within the next phase of the SSP. The design consists of three components: (i) *sensor nodes* acting as individual entities scattered throughout the region to sense the environment; (ii) a *base station* that collects the data obtained from the sensors; and (iii) a *database* to store the data collected from the networks.

In the Swarm Sensing project, a number of fixed sensor nodes (such as hive and the weather station) will be deployed within the region of interest to collect environmental data in a specific interval (e.g. 5 minutes, 15 minutes, or 1 hour). Nevertheless, a *spatial sampling* algorithm is needed to obtain a set of representative locations depending on the number of weather stations to be deployed. On the other hand, the mobile sensor nodes are being attached to insects to collect environmental data as they move within the landscape. However, while the development of micro-sensor is still being undertaken and rich data set is not available as yet, and so this dissertation also proposes a framework that generates high-resolution irregularly spaced data points to allow initial statistical data analysis.

The huge amount of data collected from the micro-sensors will be irregularly-spaced in location and time. Because of the need to have a continuous environmental variable (e.g. temperature, relative humidity, etc) over the spatial surface to obtain a clear picture of the

area under study, interpolation and extrapolation of missing values is critical to estimate the value at locations without data sampling. Therefore, this dissertation proposes a suitable spatio-temporal interpolation algorithm to accurately model the environmental situation within the RoI.

1.3 Research Objectives

The main objective of this thesis is to develop a near-to-reality Swarm Sensing Project field simulation framework and high-resolution environmental modelling within a small-scaled RoI. This work focuses on the following research components to achieve the objective:

1 To propose a *spatial sampling* algorithm to obtain representative locations for a set of static sensor nodes under the study region;

2 To propose a data-driven modelling to infer bee behaviour at the colony level;

3 To propose a Swarm Sensing field simulation framework to generate high-resolution observations (data sampling) within the region of interest from fixed (weather stations) and mobile (insects) sensor nodes.

4 To propose a spatio-temporal interpolation that suits the purpose for swarm sensing applications, based on the following factors: able to incorporate huge amount of data, low estimation error, and to create a smooth surface that accurately represents the real environmental situation.

1.4 Thesis Structure

Chapter 2 provides a summary of the current related research work in environmental sensor networks (ESN), computational bee modelling, and related interpolation techniques. Chapter 3 provides an introduction to the materials utilised in this work, followed by the optimisation algorithm for the hybrid ESN deployments (data sampling purpose) and the proposed spatio-temporal interpolation technique (high-resolution environmental modelling). Then a number of experimental simulations are used to assess the proposed method, and finally a conclusion is given in Chapter 6.

Chapter 2

Literature Review

A comprehensive overview of current technologies relating the objectives of this dissertation will be presented in this chapter. As discussed in Chapter 1 (Introduction), the ultimate goal of this work is the construction of a close-to-reality environmental model using data acquired from hybrid (static and mobile) sensor networks (Section 1.3). As such, this chapter will be partitioned into three major themes: (a) environmental sensor networks and their deployment; (b) computational insect behaviour modelling; and (c) spatial and spatio-temporal interpolation techniques. A summary discussion on how each component would contribute to this work (i.e. the swarm sensing simulation framework in Chapter 3) will be reported at the last section.

2.1 Environmental Sensor Networks (ESN)

The environmental study is an important area that provides a better understanding of natural situation and also offers a significant contribution to our society. The collected data from the sensor network is crucial for environmental managers to support decisions for an effective use of natural resources for other significant applications of benefit to our society (e.g. hazard warning services). Institutions that provide warning services often do not have enough quality data (regarding spatial and temporal coverage) that can be used to support highly accurate predictions [2–4]. Therefore, ongoing enhancement of the system's components (i.e. sensor nodes, communication engineering, power management, stability, security) is a demanding task. Hart and Martinez (2006) have provided an extensive review of ESN usage examples and they also envisioned that ESN could become a standard research tool in the near future [3].

A sensor network consists a set of inter-connected devices (Figure 1.2) that is capable of reconnaissance and surveillance for designated purposes, such as military sensing, traffic

control, industrial automation, and environmental monitoring [2, 3, 5]. It combines sensing, communicating, and computing capabilities (e.g., hardware, software, and algorithms). Such systems evolved from being passive logging systems (involving human-exhaustive effort, i.e. system maintenance and data collection) to 'intelligent' networks, where advances in cyber-infrastructure have driven the need for better sensor networks to suit different needs for designated applications or research studies. For example, enhancements in electronic miniaturisation and for wireless technologies over past decades have opened up the following unprecedented opportunities [4, 5]: (a) Making measurements at previously inaccessible and dangerous locations; (b) Improving the data quality in terms of spatial and temporal scales; (c) Obtaining unexpected observation results that led to the development of new paradigm; (d) Collaborating work between researchers across distinct professions.

An understanding of the following ESN components is required [2–6] prior to building a fit-for-purpose sensor network:

i *The purpose of the sensor network.* This includes the form of data collected and its interpretation, which has a significant effect on the entire design of the system network (e.g. communication technologies, security).

ii *Technological capabilities and the physical environment.* This relates to the deployment feasibility of sensor network, such as in mountainous or deep marine regions; and the hardware's ability to withstand hazardous situations.

iii *System standardisation and usability.* Variations in data format, hardware and software design cause difficulties in interoperability, especially when the system requires considerable amount of technical knowledge to be operated and maintained by professionals from different backgrounds.

iv *Other sensor node design goals*, including types of data (e.g. use of biotechnology), sensor integration, size, robustness, power management.

Spatial sampling methods in ESN deployments make an effort to locate sensor nodes in a way that meets desired design goals [6]. Those methods have attracted wide interest in the ESN deployments research area. The following sub-sections investigate the different types of spatial sampling and discuss the swarm sensing requirements for ESN design application.

2.1.1 Sampling with Fixed Nodes

Spatial sampling method is the effort utilising an algorithm (usually involves computational method) to deploy a fit-for-purpose sensor network that meets a set of user-predefined

requirements. It has been categorised as an NP-hard (non-deterministic polynomial-time hard) problem [6, 7], where heuristic-based approaches (mathematical programming) have been extensively used to address this problem.

Evolutionary algorithms have been utilised to optimise the sensor node placement for distinct purposes [8–12]. This technique has been widely used for optimisation mainly owing to its capability for solving multi-objective problems, where a set of near-optimal solutions (the Pareto Front, PF) will be generated as the result. In some cases, the network manager is required to possess a certain level of domain knowledge to select a particular solution from the PF [10] by considering other factors, such as the desired network's sparsity level or the feasibility of the sensor node deployments. In addition, spatial simulated annealing has also been utilised in several studies [13–17] and proven to be practical for spatial sampling purposes.

Despite the variety of optimisation techniques used for ESN deployments, the literature has reported a wide range (yet similar) of design goals. The aspects that are most-discussed and incorporated within the WSN deployment process [6]: (i) network lifetime [7, 18, 19]; (ii) connectivity [17–20]; (iii) coverage [19, 21]; (iv) relay count [19, 20]; (v) data fidelity [21, 22]; and other aspects such as cost, energy, and minimal sensor counts [7, 11, 17, 19, 21, 22].

2.1.2 Sampling with Mobile Nodes

Mobile sensor node is referring to a moving sensor node during measurements after its deployment [6]. One of many examples of mobile sensor sampling is the usage of a robot that requires continual monitoring of the environment to determine the time (when), location (where), and procedure (how) to perform the relocation. Several studies have been reported regarding the robot's positioning scheme in order to boost the network's performance with additional consideration of network coverage and travel distance [23–25]. Such systems have been widely utilised for different applications in recent wireless sensor developments, such as base station repositioning to optimise data transmission power [26], autonomous vehicle control [27], animal tracking [28].

2.2 Social Insect Modelling

This section reviews relevant topics regarding social insect behaviours (with particular an emphasis on *Apis mellifera*, the European honeybee). The following review includes honey bee behaviour (i.e. bee's foraging roles) and computational bee modelling applications. These components are crucial because one of the major contributions in this dissertation is to propose a swarm sensing field simulation, using insects as the mobile sensor nodes (bee in this case), to generate a highly dense data points in the region under study.

2.2.1 Bee Foraging Roles

The development of honey bee foraging can also be broken into several distinct roles depending on the bee's age and its knowledge of profitable food resources [29]. Initially, novice bees make several orientation flights to familiarise themselves with the landscape around the hive [30]. Such effort is crucial for successful homing after future foraging activities. Over time, bees may start to spontaneously search for food sources. These bees can be seen as 'scouts' as they are effectively naive to the availability and/or proximity of food resources in their foraging range [31]. Alternatively, the same bee whilst in the hive may observe a waggle dance, and utilise the positional information gleaned from the dance to locate food resources located by other scouts, and there-by become a 'recruit' [32–34].

Once the food source is found, the bee shifts to become an 'exploiter'. It remembers the exact location of the source, flying backwards and forwards between the hive and the source to retrieve nectar, pollen, and/or water until the source becomes exhausted or the hive's needs change. During this time, the bee may perform the waggle dance to inform other bees of the source's profitability and location [35].

Bees will cease exploiting the resource once it is exhausted, and may either become *resting* foragers or perform reconnaissance flights (as an 'inspector') to examine whether the source has replenished and, if so, will start exploiting it again [36]. Alternatively, they may start to scout for unknown resources or observe waggle dances to again become a recruit. As a bee becomes more experienced, it tends to retain a constant trip duration but with much faster speed and larger area coverage [30].

Although most of the foragers are collecting nectar and pollen, some bees are exclusively specialised on water collection (acting as a 'water carrier') with shorter and more constant flight duration [37]. Water is necessary for hive maintenance in terms of honey curing and control of hive temperature and relative humidity.

2.2.2 Computational Modelling of Bee Behaviour

Numerous attempts have been made to model both in-hive and out-of-hive bee behaviours. Pirk *et al.* have provided a comprehensive review of statistical guidelines for in-hive bee experimental design [38]. They have concluded that parametric testing (i.e. regression, normal distribution) and multivariate analysis (i.e. principal component analysis) are suitable methods for data analysis purposes (statistical inference) in bee research.

A recent study proposed an algorithm to post-process bee experimental detection data, to reconstruct bee foraging behaviour, collected from RFID (Radio-Frequency Identification) tags with a reader installed at the hive and feeder entrances [39]. When developing this 'Track-a-forager' algorithm, two major issues needed to be addressed prior to data analysis: (a) *rapid-succession* scans where successive detections were recorded in a very short time interval; and (b) *missed readings* mainly caused by the hardware limitation including small tag sizes leading to low detection ranges. Despite these issues, the 'Track-a-Forager' program has been successfully developed and is appropriate for analysing foraging bee behaviour based on the assumption that the minimal foraging duration is five minutes.

Vries *et al.* modelled individual bee foraging behaviour by considering the in-hive communication between foragers to investigate the parameters that influence foraging behaviour in a bee colony [31]. The motivation for this simulation was the fact that food profitability information encoded in the bee's waggle dance has a substantial influence on foraging behaviour at colony level [36]. It is also found that the food source positional information and the probability of a bee abandoning an exhausted source are important factors when developing an accurate foraging model. For this reason, Granovskiy *et al.* used both field experiments and a mathematical model to demonstrate the influence of the waggle dance on bee foraging behaviours at a colony level on both short and long timescales [36].

An out-of-hive foraging simulation has also been proposed by Adeva to model bee foraging behaviours during food resource depletion and replenishment [40]. This model also incorporated distinct weather parameters in its final version. The simulations generated lacked benchmark behavioural data, making it impossible to validate the model and therefore validating the accuracy of the system.

Later, Becher *et al.* modelled both in-hive and out-of-hive bee foraging behaviours when bees were exposed to different stressors including the parasite and disease vector, the Varroa mite and pesticides [40]. If accurate, such systems would be extremely useful in assisting beekeepers and biosecurity managers to understand and predict Varroa's impact on bee colonies and dispersal, thereby enabling the development of effective Varroa management systems and policy advice.

Application of these technologies to the problem of colony collapse disorder will not only help to further the development and improve the application of technology, but will also serve to assist in solving the crisis facing bee populations. Use of the technologies in this setting allows for field testing of ESNs and improvement of these devices. The data gathered during these experiments will also allow for monitoring of honey bee colony health, providing baseline data, and colony health can also be manipulated using commonly applied chemicals, for instance, to test their effects on bees in the field, furthering research into bee health and disease investigations.

2.3 Interpolation Algorithms

Interpolation is a method that is used to estimate any unknown values within known points. For example, given a set of data points/values $V = \{v_1, v_2, v_3, \cdots, v_i, \cdots, v_N\}$ at N-locations based on a function $f(x)$, for which we do not have an analytical expression of this function. The aim of an interpolation method is to construct the original function f to estimate any arbitrary spatial locations.

Interpolation is a crucial process where we want to get the information about meaningful values in the area of interest and has been widely used in many disciplines. It plays a critical role in the environmental sciences, because of the fact that the environmental analyst requires spatially continuous data over the area of interest to make valid and confident judgements.

The list below summarises the theoretical features associated with interpolation techniques [41]. Understanding these characteristics is crucial for the environmental manager because of the fact that there is no single interpolator that suits every situation. A thorough investigation is required to select the 'best suited' method for a particular purpose.

- *Global versus local.* Global methods utilise the entire observation sample for estimation to capture the general trend. Local methods, on the other hand, only considers the samples within a specified distance from the point to be estimated and so are capable of capturing the local variance [42].

- *Exactness.* This element is determined by whether or not the method will estimate the same value at sampled locations. Some examples of an 'exact interpolator' are the nearest neighbour (NN) and triangulation irregular network (TIN) methods [41]. In addition, some statistical error measurements (e.g. leave-one-out cross-validation) adopted the 'exactness' feature to determine the quality of 'inexact interpolators' [42].

- *Deterministic versus stochastic.* A stochastic method provides an estimation of the measurement's error, while a deterministic method does not.

- *Gradual versus abrupt.* This defines the *smoothness* of the estimated continuous surface produced. Gradual methods generate smooth and gradual changes between the sample observations; In contrast, an 'abrupt' technique (i.e. nearest neighbour interpolation) will produce sharp boundaries in the interpolated surface.

- *Convex versus non-convex.* A 'convex' interpolator will estimate values between the minimum and maximum values of the samples (i.e. $min \leq estimate \leq max$). A 'non-convex' method, on the other hand, might produce estimations that are lower and greater than the minimum and the maximum values, respectively. An example of such occurrences is in Kriging where some samples could produce negative weighting values resulting from the 'screen effect' [43].

- *Univariate versus multivariate.* 'Univariate' methods use only one primary variable (i.e. values at the sampled locations) for estimation, some examples are inverse distance weighting, simple kriging, ordinary kriging, and triangle irregular network. 'Multivariate' interpolators utilise more than one variable within the process, for instance, ordinary cokriging (OCK) and kriging with external drift (KED).

There has been an increased demand for interpolation techniques to incorporate spatio-temporal data. Many efforts have been initiated to address this research problem – the development of Spatio-Temporal Interpolation (STI) algorithm.These adopt either [44, 45]: (a) an *extension* approach – It converts the 'temporal' dimension into spatial-distance. On other words, it extends the STI problem into a higher spatial interpolation problem; or (b) a *reduction* approach – Such a method reduces the STI technique to a regular interpolation problem in a way that it estimates the values using spatial-only interpolation method, and then applies a time-function to incorporate the 'temporal' element within the estimation.

The following sub-sections will review some widely-used techniques providing, for each method, a description of both spatial-only and STI algorithms. Finally, a summary of applications and related research using these techniques will be presented.

2.3.1 Inverse Distance Weighting (IDW)

Inverse Distance Weighting (IDW) is one of the most widely used interpolation techniques, proposed by Shepard in 1968 [46, 47]. This technique is categorised as a deterministic method and is based on the assumption that the value to be interpolated is likely to be more similar to the nearer observed values than to those at a greater distance. This technique is expressed in the following:

$$\hat{f}(x,y) = \sum_{i=1}^{N} \lambda(d_{s_i})_i \cdot v_i \tag{2.1}$$

$$\lambda(d_{s_i})_i = \frac{d_{s_i}^{-u_s}}{\sum_{i=1}^{N} d_{s_i}^{-u_s}} \tag{2.2}$$

$$d_{s_i} = \sqrt{(x_i - x)^2 + (y_i - y)^2} \tag{2.3}$$

where $\hat{f}(x,y)$ indicates that we are estimating the value at location x,y; $\lambda(d_{s_i})_i$ is the weighting mechanism; d_{s_i} is the spatial distance (2-dimensional Euclidean distance) between the point to be interpolated (x,y in this case) and the known data point (v_i); u_s is a user-defined parameter that is used to adjust the diminishing strength in relation to increased spatial distance; and, N is the total number of known points. If the configuration u_s=2 is applied, this IDW becomes the Inverse Distance Squared technique [48]. Yet, the parameter u_s need not always be to two, and can be adjusted to improve performance [49]. The complexity of conventional IDW is $O(N)$ which can be seen from Equation 2.1.

The so-called 'zero distance problem' has been discussed by de Mesnard [50]. For the case where location to be interpolated is exactly the same as one of the reference points (i.e. $d_{s_i} = 0$), Shepard [46] does not interpolate that particular location because we already have full knowledge at that point – the *discrete* case. Unfortunately, such implementations may not be realistic when the mean within a particular area (suburb, city, state, country) is being estimated (i.e. $d_{s_i} \to 0$) and, for these situations, utilising the *continuous* case is more satisfactory.

Improved IDW

IDW is based on the notion that nearer data points will have more influence compared to those further away, and so including every data point throughout the map to interpolate a single point is unnecessary. This is because, as the distance is further away, the particular point has very little influence on the original value. An illustration of the improved version of IDW can be seen in Figure 2.1, and the parameter R can be optimised to improve the quality of IDW [49]. The improved IDW comes with two major advantages in terms of computational efficiency: (a) The processing time does not increase as the number of known points increases; (b) We can further improve the performance by applying the kd-tree data structure algorithm, reducing the computational time from $O(N)$ to $O(\log N)$ [51]. This feature is crucial because the basic computational time required for the STI algorithm is $O(T \times N)$, instead of $O(N)$, where T is the number of user-defined window lengths to be included in the interpolation process.

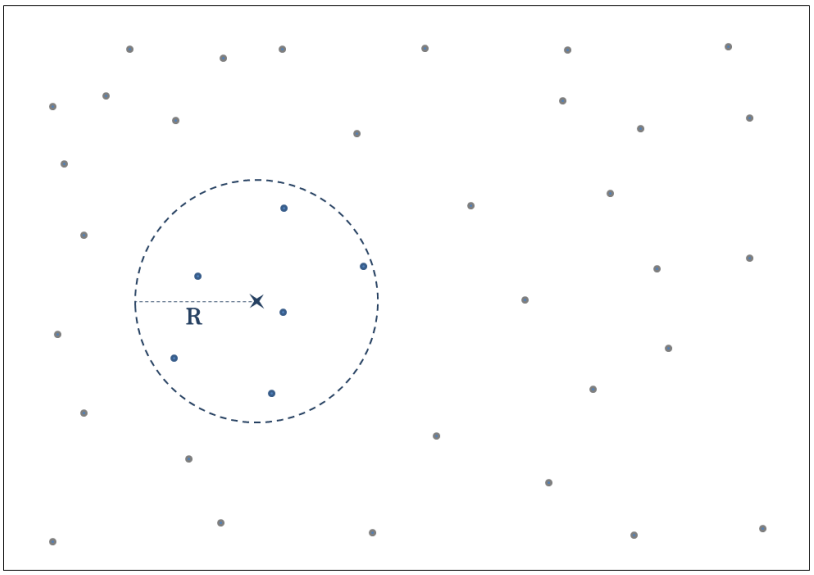


Fig. 2.1 A demonstration of the improved IDW. The dots are the sample data points, and the '×' is the point location to be interpolated. R is a user-defined radius parameter that indicates the farthest distance to be included from the point ×. On this case, only a total of six sample points will have an influence on the interpolation process. However, no empirical approach has been developed to obtain the optimal value for the parameter R.

STI - Extension approach

The extension approach to the IDW's STI method (2-D Space and 1-D Time problem) was described by Li and co-authors in 2014 [51]:

$$\hat{f}(t,x,y) = \sum_{t_{start}}^{st_{end}} \sum_{i=1}^{N} \lambda(d_{st_i})_{st,i} \cdot v_{t,i} \tag{2.4}$$

$$\lambda(d_{st_i})_{st,i} = \frac{d_{st_i}^{-u_{st}}}{\sum_i^N d_{st_i}^{-u_{st}}} \tag{2.5}$$

$$d_{st_i} = \sqrt{(x_i - x)^2 + (y_i - y)^2 + c^2(t_i - t)^2} \tag{2.6}$$

where d_{st_i} is the spatio-temporal distance between the measured (x_i, y_i, ct_i) and unmeasured (x, y, ct) location point; u_{st} is the spatio-temporal diminishing strength as distance d_{st_i} increases; and, c is the user-defined temporal factor that converts the time unit to a spatial distance unit; However, there is still no empirical information on how to justify choice of the temporal factor (c), and a naive choice does not yield optimal results [51].

The extension of this approach to the 3-D space and 1-D time problem can be expressed in a slight variation based on Equation 2.6, so that it becomes:

$$d_{st_i} = \sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2 + c^2(t_i - t)^2} \tag{2.7}$$

where z is the 3^{rd} spatial dimension in which always seen as the altitude (surface height).

STI - Reduction approach (ST Product Method)

This method was proposed by Li *et al.* and is constructed in the following way [45]:

$$\hat{f}(x,y,t) = \sum_{i=1}^{N} \lambda(d_{s_i})_{s,i} \cdot \hat{f}(t) \tag{2.8}$$

$$\hat{f}(t) = \frac{t_{i2} - t}{t_{i2} - t_{i1}} v_{i1} + \frac{t - t_{i1}}{t_{i2} - t_{i1}} v_{i2} \tag{2.9}$$

where the spatial weighting $\lambda(d_{s_i})_{s,i}$ is equivalent to Equation 2.2; $\hat{f}(t)$ is the estimated value at time t; t_{i1} and t_{i2} are the first (previous) and second (next) time indices, and similarly, v_{i1} and v_{i2} are the first (previous) and second (next) value at corresponding time t.

It is important to note that such an algorithm (Equation 2.9) relies on the assumption that values at the same location (v_{i1} and v_{i2}) but at different times (t_{i1} and t_{i2}) are provided. Nevertheless, one of the assumptions of this dissertation is that the data are not necessarily collected in a finely-gridded manner (in both the spatial and temporal dimensions). Due to the fact that this algorithm does not meet the assumptions of this thesis, it will not be considered and applied to the simulation in this work.

2.3.2 Kriging

The Kriging interpolation technique, also called the Kriging esimator, is categorised as a geo-statistical method because it takes the spatial patterns and the uncertainty of the surrounding surface into account during the interpolation process [42]. An observation $Z(s,t)$ can be seen as a combination of a space-time mean component $m(s,t)$ and a stochastic residual component $Y(s,t)$, written as:

$$Z(s,t) = m(s,t) + Y(s,t) \tag{2.10}$$

with m representing a deterministic 'global' trend and Y the corresponding zero-mean noisy residual. The following sub-sections will discuss the process of Kriging (characterising trend

and residual), and an example of the Ordinary Kriging (OK) calculation will be provided at the end.

Trend Characterisation

The behaviour of an observation variable is different at different spatial and temporal scales and so can be characterised using a combination of linear models as follows:

$$m(s,t) = \sum_{i=0}^{p} \beta_i f_i(s,t) \tag{2.11}$$

where β_i is an unknown regression coefficient; f_i represents the covariates that must be known exhaustively over the space-time domain; and p is the number of covariates.

Stochastic Residual Modelling – Variogram

A *semi-variance* is generated to show how much a location point is related to its neighbour points within a particular distance (called the *lag*) by using the following equations for the spatial-only and the spatio-temporal cases:

$$\hat{\gamma}_s = \frac{1}{2} E\left[Y(s) - Y(s+h_s)\right]^2 \tag{2.12}$$

$$\hat{\gamma}_{st} = \frac{1}{2} E\left[Y(s,t) - Y(s+h_s,t+h_t)\right]^2 \tag{2.13}$$

where E denotes mathematical expectation; $Y(s)$ and $Y(s+h_s)$ is the residual value at spatial location s and $s+h_s$ (separated by spatial lag distance h_s) respectively. The semi-variance $\hat{\gamma}(h_s)$ is plotted against h_s, and needs to be fitted in order to create the so-called *variogram model* for the estimator process in the later stage.

Some widely-used spatio-temporal stochastic *semi-variance* models are briefly discussed below:

(a) Sum-metric model. This model is based on the assumption that the three components (spatial, temporal, and spatio-temporal) are mutually independent [52]:

$$\gamma_{st}(h_s,h_t) = \gamma_s(h_s) + \gamma_t(h_t) + \gamma_{st}\left(\sqrt{h_s^2 + (\alpha \times h_t)^2}\right) \tag{2.14}$$

where h_t is the temporal distance lag; and α is the spatio-temporal anisotropy ratio that converts the unit of temporal separation (h_t) into a spatial distance (h_s).

(b) Product-sum model. Proposed by de Iaco [53] in the form of:

$$\gamma_{st}(h_s,h_t) = \gamma_{st}(h_s,0) + \gamma_{st}(0,h_t) + k\gamma_{st}(h_s,0)\gamma_{st}(0,h_t) \tag{2.15}$$

with

$$k = \frac{sill\gamma_{st}(h_s,0) + sill\gamma_{st}(0,h_t) - sill\gamma_{st}(h_s,h_t)}{sill\gamma_{st}(h_s,0)sill\gamma_{st}(0,h_t)} \tag{2.16}$$

where $sill\gamma_{st}(h_s,h_t)$ denotes the 'global' sill estimated by 'eye-fit' after plotting the sample variogram surface ($\gamma_{st}(h_s,h_t)$) or by fitting to minimise the least-squares error of Equation 2.15 [53]. Also, the following must be met in order to satisfy the admissibility condition for k:

$$0 < k \leq 1/\max\{sill\gamma_{st}(h_s,0), sill\gamma_{st}(0,h_t)\} \tag{2.17}$$

Ordinary Kriging (OK) Estimation

The weighting mechanism for the OK estimator is formulated by solving the equation:

$$A^{-1} \cdot b = \begin{bmatrix} \lambda \\ \phi \end{bmatrix}, \quad A = variogram\_model(D_{N,N}) \tag{2.18}$$

$$D_{N,N} = \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,N} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N,1} & d_{N,2} & \cdots & d_{N,N} \end{pmatrix} \tag{2.19}$$

where $D_{N,N}$ is a $N \times N$ distance matrix between the known points, and A is the matrix holding the values after applying the *variogram model* to $D_{N,N}$; λ represents the weights between the location to be interpolated and the known points ($\sum \lambda = 1$); and ϕ is the Lagrange multiplier. Finally, the estimation variance ($\hat{\sigma}_e^2$) of the OK estimator can be calculated using:

$$\hat{\sigma}_e^2 = \sum_{i=1}^{n} \lambda_i \gamma(d_{s_i}) + \phi \tag{2.20}$$

A detailed description and a step-by-step example calculation of the OK estimator is described in greater detail at [42].

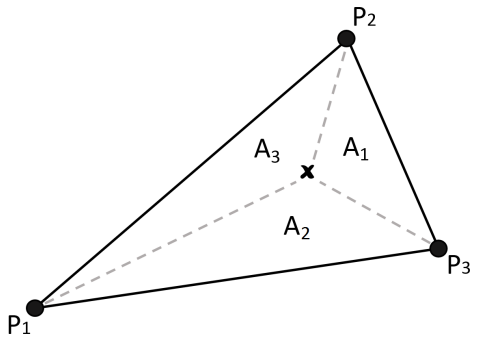


Fig. 2.2 An illustration of the TIN interpolation technique for point x that lies within a triangle formed by points P_1, P_2, and P_3 (holding values of v_1, v_2, and v_3 respectively). The weight for each point is calculated based on the corresponding area; for instance, P_1 has the weight $\frac{A_1}{A}$ (where $A = \sum_{i=1}^{N=3} A_i$) and so on.

2.3.3 Shape Functions (SF)

A Shape Function (SF) based spatial interpolation technique employs a Triangular Irregular Network (TIN) for 2-D spatial-only interpolation purposes [45]. TIN is a digital means of representing surface morphology and has been extensively used in the GIS (Geographic Information System) community. It is produced by connecting edges between vertices that eventually form a network of triangles, and is normally constructed using the Delaunay triangulation algorithm.

This surface analysis technique can be further extended as a linear approximation interpolation algorithm as proposed by Peuker and co-workers in 1978 for digital elevation modelling [54]. It was described in the work done by Li in 2003, which uses area divisions for the weighting mechanism [45] [Figure 2.2]. As it is based on triangle meshes, the total number of included observed points is 3. It is in the form of:

$$\hat{f}(x,y) = \lambda_1 v_1 + \lambda_2 v_2 + \lambda_3 v_3, \qquad \lambda_i = \frac{A_i}{A} \tag{2.21}$$

where A is the area of the entire triangle according to:

$$\begin{aligned}
\frac{A_1}{A} &= N_1(x,y) = \frac{[(x_2y_3 - x_3y_2) + x(y_2 - y_3) + y(x_3 - x_2)]}{2A} \\
\frac{A_2}{A} &= N_2(x,y) = \frac{[(x_3y_1 - x_1y_3) + x(y_3 - y_1) + y(x_1 - x_3)]}{2A} \\
\frac{A_3}{A} &= N_3(x,y) = \frac{[(x_1y_2 - x_2y_1) + x(y_1 - y_2) + y(x_2 - x_1)]}{2A}
\end{aligned} \tag{2.22}$$

$$A = \frac{1}{2}\, det \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \tag{2.23}$$

where A_i is the i^{th} sub-triangle's area formed by the point to be interpolated (x in Figure 2.2); x_i and y_i is the i^{th} node (i.e. known point at location x and y).

STI - Extension approach

Li extended the SF-based STI technique to become a 3-D triangular object (2-D space and 1-D time) called a tetrahedra mesh [45]. Such a function can be generated automatically using the Delaunay refinement algorithm [55] and its improvement using swapping and smoothing [56]. Similar to the spatial-only SF interpolation technique described previously, Eq. 2.21 and Eq. 2.23 now become:

$$\hat{f}(x,y,t) = \lambda_1 v_1 + \lambda_2 v_2 + \lambda_3 v_3 + \lambda_4 v_4, \qquad \lambda_i = \frac{V_i}{V} \tag{2.24}$$

$$V = \frac{1}{6}\, det \begin{bmatrix} 1 & x_1 & y_1 & t_1 \\ 1 & x_2 & y_2 & t_2 \\ 1 & x_3 & y_3 & t_3 \\ 1 & x_4 & y_4 & t_4 \end{bmatrix} \tag{2.25}$$

where V_i is the i^{th} sub-component volume based on the entire tetrahedra volume V; and t_i is the i^{th} node at time t.

STI - Reduction approach

The reduction-based STI method for SF is similar to the one demonstrated in IDW's STI reduction approach, which is based on Equation 2.8 and Equation 2.9. Consequently, by

applying Equation 2.22 to Equation 2.8, the final form of SF's reduction based STI technique can be re-written as [45]:

$$\begin{aligned}\hat{f}(x,y,t) = \quad & N_1(x,y)\left[\frac{t_2-t}{t_2-t_1}v_{1,1}+\frac{t\_t_1}{t_2-t_1}v_{1,2}\right] \\ & +N_2(x,y)\left[\frac{t_2-t}{t_2-t_1}v_{2,1}+\frac{t\_t_1}{t_2-t_1}v_{2,2}\right] \\ & +N_3(x,y)\left[\frac{t_2-t}{t_2-t_1}v_{3,1}+\frac{t\_t_1}{t_2-t_1}v_{3,2}\right] \\ = \quad & \frac{t_2-t}{t_2-t_1}\left[N_1(x,y)v_{1,1}+N_2(x,y)v_{2,1}+N_3(x,y)v_{3,1}\right] \\ & +\frac{t-t_1}{t_2-t_1}\left[N_1(x,y)v_{1,2}+N_2(x,y)v_{2,2}+N_3(x,y)v_{3,2}\right]\end{aligned} \tag{2.26}$$

where t_1 and t_2 are the first (previous) and second (next) time index before / after the time (t) to be interpolated; $v_{i,j}$ is the value at node i and time j where $j = \{1,2\}$ (note that $j = 1$ and $j = 2$ is equivalent to t_1 and t_2 respectively).

2.3.4 Applications and Related Work

This subsection reviews related work involving spatial-only and spatio-temporal interpolation techniques and applications. For the purpose of this dissertation, the works reviewed mainly focus on the environmental sciences area.

Spatial-only interpolators

Inverse Distance Weighting (IDW) is one of the earliest deterministic interpolation techniques to encompass simplicity and effectiveness during the estimation and interpretation process. It has been utilised to identify trends and variability in the mean at unsampled locations for particular climate variables [57]. The continuous surface generated is critical to assist stakeholders and managers to identify the risks and vulnerabilities for better decision making. Using the same approach, Chen and Liu (2012) proposed that the spatial distance-decay parameter u_s in Equation 2.2 could highly influence the accuracy of this method so should be chosen carefully [49]. They also adapted the 'improved' IDW technique (Section 2.3.1) to limit the search area and concluded that the number of weather stations also affect the performance of the technique.

Owing to the simplicity of IDW in terms of computational efficiency and interpretation, some researchers have further enhanced the performance of pure IDW. For example, in 2008, Lu and Wong proposed an 'Adaptive' IDW (AIDW) that could accommodate the

samples' sparsity by varying the constant distance-decay parameter (u). They believed that the parameter u should be lower in a less-dense area [58]. Overall, their experimental results show that Kriging > AIDW > IDW (where '>' denotes 'better than'), despite the case where AIDW > Kriging in one of their empirical studies due to high spatial heterogeneity (i.e. unable to effectively model variogram functions). It is noted that AIDW surpasses pure IDW when IDW yields an acceptable outcome.

One of the main deficiencies of IDW is the fact that it is unable to provide a confidence interval for the estimation. Joseph and Kang (2011) addressed this limitation issue by developing a Regression-based IDW (RIDW) algorithm so that a Confidence Interval (CI) error measurement for IDW estimation is possible [59]. They have shown that RIDW has a similar prediction accuracy to Kriging and, interestingly, also indicated that the CI for RIDW is much better than the Kriging's CI.

Starting from the early 20^{th} century, a widely used application of the IDW-based method is to utilise historical data to generate regularly-spaced gridded environmental data sets at different resolution levels [60–62]. Despite using a purely distance weighting mechanism (Equation 2.2), the authors employed the so-called Angular IDW (AIDW) algorithm that incorporates directional isolation between the sample points to update the weights. Caesar and Alexander (2006) suggested that the distance-decay parameter $u_s = 4$ (as in Equation 2.2) is favourable in order to compromise for an acceptable statistical error and helping to reduce spatial smoothing [61]. Other spatial interpolation methods such as isolation and combination between Thin-Plate Splines (TPS) and Kriging-based methods (Indicator Kriging, Universal Kriging, and Kriging with external drift) were also utilised for spatially continuous climate data reconstruction for Australia [63] and for Europe [64].

A number of other spatial interpolation methods have been compared to assess the suitability of different techniques for mapping rainfall spatial variability. Kriging-based methods provide the most consistent results, while spline and trend surface fitting (using polynomial regression) performed poorly in most cases under one empirical study [65]. Plouffe *et al.* utilised rainfall data for different months (May and September) and suggested that Bayesian Kriging and splines provide good estimation at low and high rainfall, respectively [66]. It is concluded that there is no 'optimal' method for this purpose and the interpolator's performance depends on the setting and the characteristics of a specific data set.

Another application of interpolation algorithms is to characterise soil properties over a region of interest. Two widely used interpolation techniques have been investigated for this purpose by Gotway *et al.*, namely OK and IDW with different distance-decay parameter values [67]. They suggested that the IDW's distance-decay parameter should be altered based on the coefficient of variation (CV) of the data set, but OK still provides the most

accurate prediction with the cost of heavy computational effort. Meul and Meirvenne studied the stationarity component of soil properties using geostatistical analysis [68]. They compared the performance of Kriging-based methods under different forms of nonstationarity – Universal Kriging (UK), Simple Kriging with varying local means (SKlm) and Ordinary CoKriging (OCK). They reported that different methods performed better under distinct circumstances: (i) OCK is preferable when there is a high correlation between the primary and secondary variables; and (ii) UK is best when local nonstationarity is present. In addition, they have found that Kriging yields the highest precision when utilising a combination of UK and OCK. In spite of the aforementioned interpolation applications for soil properties, Schloeder *et al.* (2001) doubted the accuracy of interpolators because of the following factors: spatially independent data, limited data, sample spacing, extreme values, and erratic behaviour [69].

As well as the application of deterministic and geostatistic mechanisms for interpolators, some authors have proposed Machine Learning (ML) based interpolation algorithms. For instance, Sun *et al.* not only compared the performance of various Kriging-based techniques (simple kriging, OK, UK) and IDW, but also included the Radial Basis Function (RBF). Their result indicates that ML-based methods (RBF in this case) may not necessarily be better than the conventional interpolators: SK > IDW > RBF > OK > UK [70]. In 2013, Matos *et al.* have also proposed several ML-based techniques in addition to RBF, they are: Support Vector Regression (SVR) and Least Squares Support Vector Regression (LS-SVR). They concluded that ML-based methods could produce superior outcome under certain condition [70, 71] in which has been shown in [72]. However, one should also consider the excessive amount of time required for the process of ML-based estimator.

Lastly, a hybrid method has been proposed and compared by Sanabria *et al.* [73] for wildfire risk assessment in Australia. They have investigated IDW (non-geostatistical), OK (geostatistical), Random Forest (RF, machine learning), and also a combination of methods – RFIDW and RFOK. They concluded that the proposed hybrid method achieved better estimation than any of the isolated methods, even though RF-based method is more computationally demanding.

Spatio-temporal Interpolators

Li and Revesz (2003) compared the three most widely used Spatio-Temporal Interpolation (STI) techniques. They also made comparisons between the different STI approaches: (a) *Extension* – TIN, Kriging and IDW; (b) *Reduction* – TIN and IDW [74]. They reported that TIN is the most suitable method for house pricing interpolation regardless of the type of STI method used. In this case, TIN could be acceptable because the data type of house

pricing varies in a way that does not necessarily involve gradual changes over the surface (as environmental variables do). It is interesting to see that, reduction-based IDW surpasses Kriging's performance. The extension-based IDW exhibited the worst performance probably because the space-time interaction component has not been extensively studied (i.e. the anisotropy ratio between space and time). Li *et al.* (2011) utilised an identical extension-based TIN method to that described for an air pollution application [75]. Within the study, the authors 'scaled' the temporal data in an effort to increase the prediction accuracy; However, an 'optimal' time-scaling method for the data set has yet to be found.

A fast extension-based IDW STI algorithm was proposed by Li *et al.* [51]. They put the main focus on the computational efficiency of the algorithm and it has been successfully achieved by utilising the following practices: (i) using the 'improved' IDW technique, as discussed in Section 2.3.1, which only includes k nearest neighbours within the estimation process; (ii) using *kd-tree* data structure; and (c) using parallel programming techniques.

In the case of geostatistical STI algorithms, Kilibarda *et al.* (2014) have successfully modelled global temperature data at high resolution (e.g., $1km^2$) using STI regression-kriging and a sum-metric variogram structure [76]. The only disadvantage of such methods is that the 'optimal' estimation of the anisotropy ratio that converts the temporal unit to a spatial-distance unit has yet to be found [51]. An alternative is the product-sum variogram model, which has been utilised by Zeng *et al.* (2012) within STI OK to estimate the carbon dioxide distribution in China [77].

2.4 Summary of the Literature

This section provides a discussion on how each of the previously reviewed components would contribute to the research work in this dissertation.

2.4.1 Hybrid ESN for the Swarm Sensing Project (SSP)

As mentioned previously, since the development of micro-sensor is still ongoing, we are unable at this stage to deploy it on insects to make observations as they move throughout the landscape. This has stimulated the need to develop a near-to-reality hybrid ESN field simulation framework to collect highly-dense environmental data, using: (a) static sensor nodes, acting as weather stations, that make measurements on an hourly basis; and (b) mobile sensor nodes, such as micro-sensors being deployed on bees, that sense the environmental conditions that they are experiencing as they move.

2.4.2 Computational bee modelling

The current literature for bee modelling mostly addresses the influence of individual bee behaviour, with the help of 'waggle dance' for in-hive communication, on colony-level foraging behaviour. For example, the encoded information within a bee's dance can be used to assess the profitability of a single food source and to determine whether it is more beneficial to keep exploiting it or to abandon the exhausted source [31, 36, 78].

Despite the extensive research done, however, there is little information that could be usefully incorporated into the SSP's field simulation. For example, the model's ability so that individual bees could sense and collect the environmental situation data (i.e. data sampling) as they fly through the landscape. Another issue is the feasibility of simulating an artificial bee behaviours throughout the day without any empirical evidence on which to base the simulation. To illustrate, assuming no domain knowledge of bee behaviour, one could ask: at what time of the day is a bee most active? and, what is the probability of a bee foraging in the morning, noon, or late at night?

This work proposes a data-driven statistical modelling framework to address these issues so that simulating an artificial bee is possible.

2.4.3 Interpolation algorithm

It is widely recognised that there is no 'optimal' interpolation technique that is perfectly suitable for all purposes [42, 65, 66, 71] and the performance of interpolators varies depending on (and not limited to): sampling density, sampling pattern, spatial structure of primary variable, and surface type [41, 48, 79]. Therefore, a careful investigation is required in order to select the most appropriate algorithm for a certain objective [42]. Nevertheless, it is also noted that some subjective aspects regarding the selection of interpolation method are also significant, such as: (i) the required level of computational efficiency in the case of extremely large data set; (ii) the applicability of the interpolator for a specific task – it is unrealistic to map temperature data using a triangular irregular network algorithm because it produces abrupt surface discontinuities; and (iii) the availability of particular information required prior to the interpolation process – Simple Kriging requires a known mean of the region under study.

For the purpose of this dissertation, based on the fact that observations made from a swarm of sensor nodes will generate a huge amount of data over a small landscape, computational efficiency becomes a crucial factor for the spatio-temporal interpolation algorithm development. The statistical error of the method will also be carefully considered. Finally, given that the ultimate purpose of the interpolation algorithm being developed for the

high-resolution environmental modelling over the region under study, then the characteristics of the algorithm output must match those of the real environmental situation being modelled. For instance, it is expected that the temperature over a small region would vary gradually and that a smooth surface would be obtained.

The Inverse Distance Weighting (IDW) interpolation method has been recognised as a computationally efficient technique that provides acceptable results [51]. Given that IDW performs well in a highly dense network [49], an 'IDW-like' method would be a promising technique to be considered and implemented in my research project. The following main challenges still remain and will be investigated in this dissertation: What statistical measures can be used to accurately represent the space-time relationships to be utilised within the spatio-temporal interpolation technique in a computationally efficient manner?

Chapter 3

Methodology

The main objective of this dissertation is to propose and to develop a computational field simulation framework and environmental modelling application for swarm sensing application. Figure 3.1 depicts the Region of Interest (RoI) on which the field simulation will be based the South Esk region, Tasmania. To begin with, Table 3.1 presents summary attributes based on different sensor types to be deployed.

Fig. 3.1 Map of the state of Tasmania. The red rectangle (in the north-eastern corner of the map) is where Tasmanian's South Esk region is located. (Image source: Google map)

The static nodes, acting as conventional large-sized weather stations, are assumed to be an individual entity that has the capability of storing each observation made (i.e. similar

Attribute	(a) Weather Station	(b) Micro-sensors
Short Description	A conventional weather station.	Micro-sensors each attached to a bee's thorax (Figure 1.1).
Sensor Size	Large	Small (sub-millimetre)
Number of Sensors	Small ($N < 10$)	Large (e.g., thousands)
Dynamics	Static	Mobile
Communication	Wired	Wireless
Energy Availability	Unconstrained	Constrained
Spatial Coverage	Small (one location point)	Large (various locations)

Table 3.1 Distinct types of sensor nodes to be deployed: (a) static weather station; and (b) mobile micro-sensors. N indicates the number of sensor nodes.

to a base station), so that network connectivity and relay capability are not issues. Regular maintenance of the system will be conducted throughout the experiment to minimise data loss during system downtime. Since it is quite costly to set up a weather station, the number of deployments is usually low. This kind of sensor node only has a very small spatial coverage but could be configured to make high frequency measurements. Considering other factors such as power consumption and data storage, it is often set up in a way that data collection occurs only within a pre-defined time interval (5m, 30m, 1h, daily, etc) under the assumption of having unconstrained energy. For the purpose of this work, weather stations are categorised as the following entities: food and/or water sources, and bee hives (Table 3.2). Note that weather stations are placed at each of these entities/locations.

Mobile nodes, on the other hand, are being deployed on insects (i.e. bees in this case) to 'sense' the environment as they fly through the landscape. A large number of measurements and spatial coverage will be achieved within the area under study. Due to limitations in the technology, energy available for individual sensor node to transfer data to the base station will be constrained. Spatial sampling (relocation) of the mobile sensor node is not the focus

of this thesis, and is impractical to do so in this case (since it would require us to manipulate the insect flight paths). Since one of the main objectives of this dissertation is to propose a framework to simulate artificial bees, bee foraging flight path data is crucial. Therefore, this work will employ the modelled bee foraging flight paths data set which have been developed by the CSIRO's Swarm Sensing team [80]. Further description of this component will be given in Section 3.1.3.

Sensor Type	Item	Deployment Strategy	Number of Nodes	Reading Frequency
Static	Food source	Manual	5	Hourly
Static	Water source	Manual	1	Hourly
Static	Weather station	Manual	4	Hourly
Static	Hive	Optimisation	5	Hourly
Mobile	Insects	Random (foragers)	20†	Minute / Seconds

† Number of nodes deployed each day, from a single hive.

Table 3.2 This table shows the configuration of the field simulation to be deployed within the area of interest (South Esk region of Tasmania, see Figure 3.1).

Table 3.2 presents the configuration of our proposed hybrid environmental sensor networks (ESN) field simulation at our Region of Interest (RoI). Based on the table, three types of stationary sensor nodes will be deployed for different purposes: (a) bee hive; (b) food and water sources for the bees; and (c) four extra weather stations located at the RoI's map corner because the environmental modelling on the later stage is an *interpolation* issue (instead of *extrapolation*). In addition, note that the table shows 20 bees will be created each day in each hive, which indicates that a total of 100 bees (5 hives $\times$ 20 bees) will be generated every day.

Lastly, the research design being undertaken in this work is as follows:

1 *Data collection*.
A number of data sources will be utilised in this work for different purposes, they are: the South Esk Hydrological Model, CSIRO's bee experimental data, an agent-based bee behavioural model developed by CSIRO, and a 'benchmark' data set for validation at the end.

2 *The 'swarm sensing' field simulation.*
Propose a framework to simulate a hybrid Environmental Sensor Network (ESN) to generate high-resolution spatio-temporal data within the RoI (i.e. data sampling).

3 *Formulating the spatio-temporal interpolation algorithm.*
Investigate related mathematical and/or geo-statistical interpolation techniques that might be relevant. Develop a 3-Dimensional (3D = 2D Spatial + 1D Temporal) spatio-temporal interpolation (STI) algorithm that meets the Swarm Sensing project requirements: computational efficiency (elapsed time); quality (low statistical error); and subjective assessment (e.g. near-to-reality environmental situation, visual preference, etc).

4 *System implementation.*
Implement the field simulation framework and the developed interpolation mathematical model into software using the Python programming language, and finally, create a realistic environmental model.

5 *Validation and analysis.*
Evaluate the algorithm that has been developed.

The following sub-sections will cover the details of materials used in this work and the framework's components as shown in Figure 3.2.

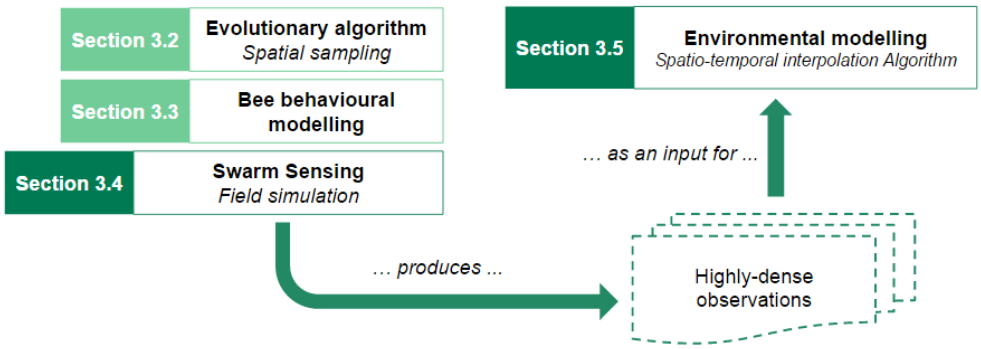


Fig. 3.2 An overview of the swarm sensing field simulation to produce highly-dense observations (Section 3.4) to be utilised as an input for the spatio-temporal interpolation algorithm (Section 3.5). Spatial sampling to optimise the hives locations (Section 3.2) and modelling bee behaviour (Section 3.3) are acting as the "supporting components" in order to accomplish the field simulation.

3.1 Material

First of all, this section will describe the list of data sets that will be employed for the purpose of this dissertation. For ease of reference in the following Chapters, the notations illustrated in Table 3.3 are used and detailed information for each data set is described in the following sub-sections.

Dataset #	Abbreviation	Data Set Name	Section #
DataSet 1	'modelled'	South Esk Hydrological Model	Section 3.1.1
DataSet 2	'experimental'	Bee Experimental Data Set	Section 3.1.2
DataSet 3	'foraging'	Bee Foraging Flight Model	Section 3.1.3
DataSet 4	'benchmark'	Benchmark Data Set	Section 3.1.4

Table 3.3 List of data set notations and abbreviations that will be used throughout the dissertation.

3.1.1 South Esk Hydrological Model

Introduction

The South Esk hydrological model developed by Commonwealth Scientific Industrial Research Organisation (CSIRO) is utilised in this work [81]. It is located at the South Esk region of Tasmania (-41.0° to -42.0° latitude and 147.0° to 148.5° longitude), Australia 3.1]. The model covers a range of distinct environmental parameters describing the RoI, which is mapped onto a 151×101 grids of a the 2-D spatial map (with 1 km^2 resolution) and is measured with high temporal resolution (1 hour interval). The data set is stored in NetCDF (Network Common Data Format) format [82].

For the purpose of this study, I utilised the data set for the year 2013 and mainly focused on four different parameters that are widely used for the environmental study: surface height (also called elevation), temperature, relative humidity, and solar radiation. It is undeniable that the data set consists of some missing data. However, by considering the integrity of the evaluation in this work (since we need 'real' data to assess the algorithm's performance), I will not attempt to 'fill in the gaps' of those missing data. Also, such an activity is outside the scope of this work.

The following subsection will provide some data analytics and visualisations so that readers will have a better understanding of the data set's overall characteristics. The color

maps that are used to plot the 2D-maps in the following sections are chosen based on the findings provided by Engelke *et al.* [83].

Analytic and Visualisation

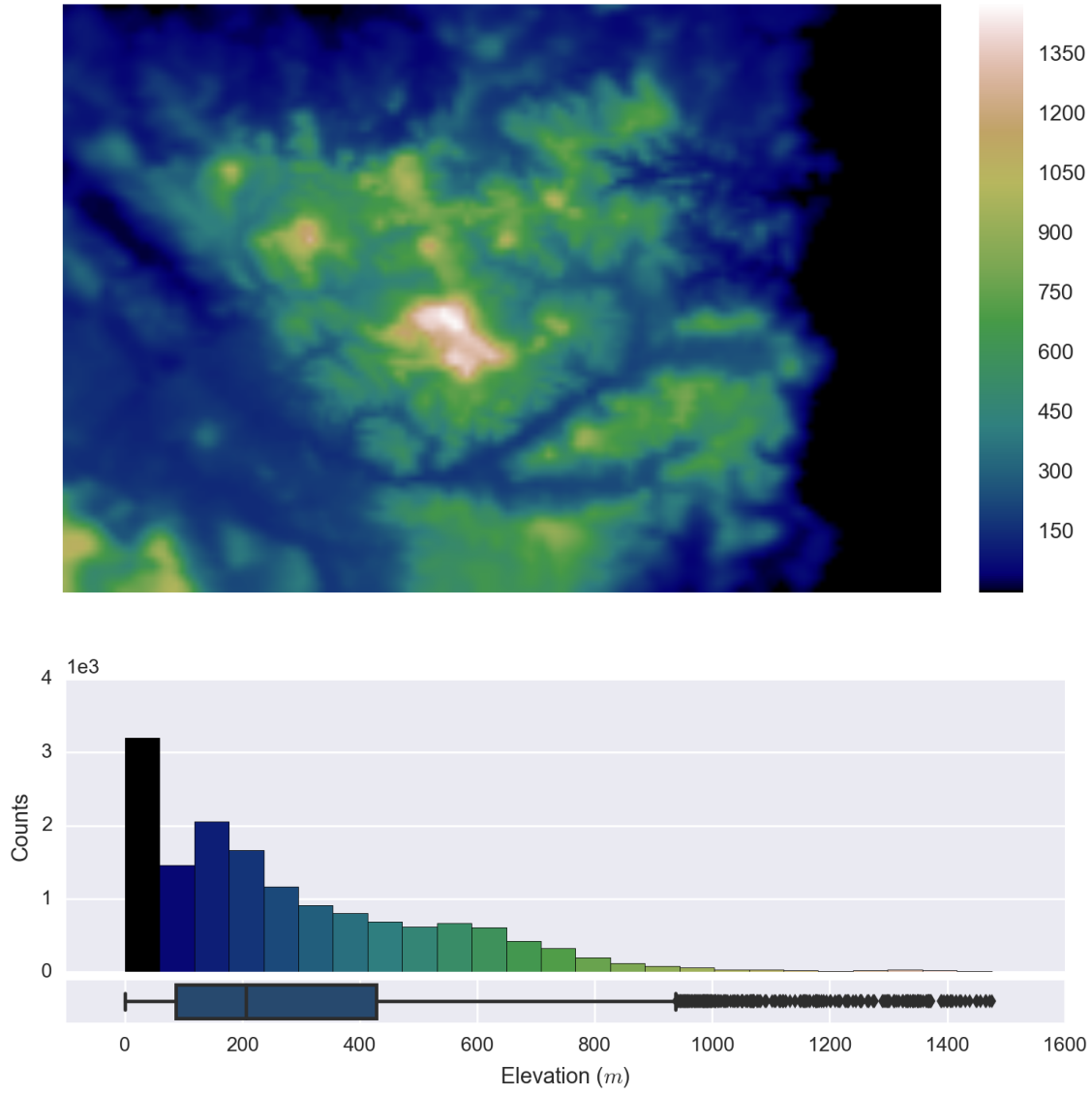


Fig. 3.3 A typical surface height data visualisation of Tasmania's South Esk Hydrological model: (top) The actual elevation data (*meters*) within the RoI; (bottom) The distribution of elevation values within the 151×101 grid.

Surface Height

Surface height data is used to describe the physical spatial attribute of the landscape that is measured from the sea level. The map consists of elevation data between 0 $meters$ (m), indicating sea level as depicted using dark blue, and 1475 m at the mountain peak which is illustrated using a brown-white colour in Figure 3.1. At the bottom of Figure 3.1 is the distribution of height data and it shows that the average height across the landscape is approximately 200 m.

Elevation data is crucial for the environmental study [84], mainly because such data is highly correlated with some environmental parameters (e.g., temperature as shown in Figure 3.5a). There are also other factors that support the notion that elevation data is a good data source to be utilised for research:

- Does not vary for a long period of time (i.e., decades); unlike other parameters such as temperature that shift within seconds.
- Can be easily obtained from different sources on the Internet.

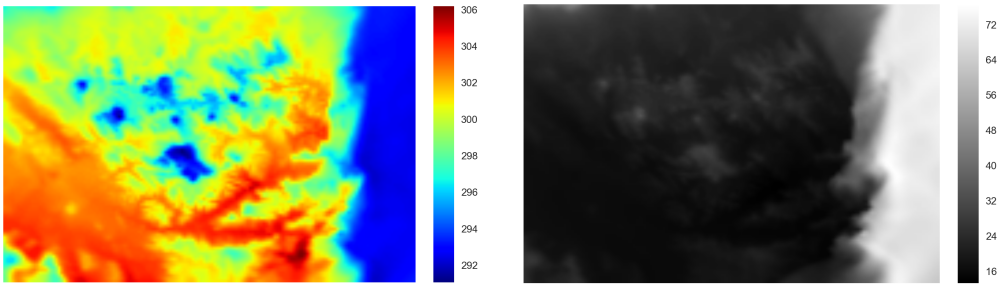
Other Parameters

The three other parameters that will be used in this work are: temperature, relative humidity, and solar radiation. Example visualisations of different environmental variables are depicted in Figure 3.4. The temperature (K) data exhibit seasonality effects with higher temperatures expected during summer (between December and February) and lower temperatures during winter (between June and August). Relative humidity data, on the other hand, is presented using a percentage ranging from 0% to 100%. Finally, similar to the temperature data, seasonal effects also exists (i.e. higher value during summer and lower value during winter) within solar radiation data.

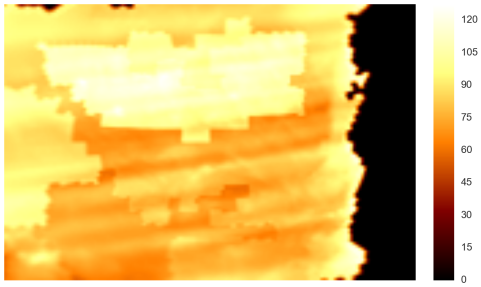
Parameter Correlation

Following, a statistical method is utilised to investigate the correlation between parameters by using the squared Pearson's correlation coefficient (r, Equation 3.27) – Coefficient of Determination (R^2). Figure 3.5 projects the correlation level between parameters after applying such a calculation.

It is observable that the two most correlated parameters are surface height and temperature, in which can be seen from Figure 3.5a and Figure 3.5b. Yet, the correlation level varies between months and the highest correlation among them occurs during April and August. Based on Figure 3.5b and Figure 3.5c, temperature and relative humidity also have fairly good correlation ($R^2 = \pm 0.4$). Finally, solar radiation is the least correlated parameter to



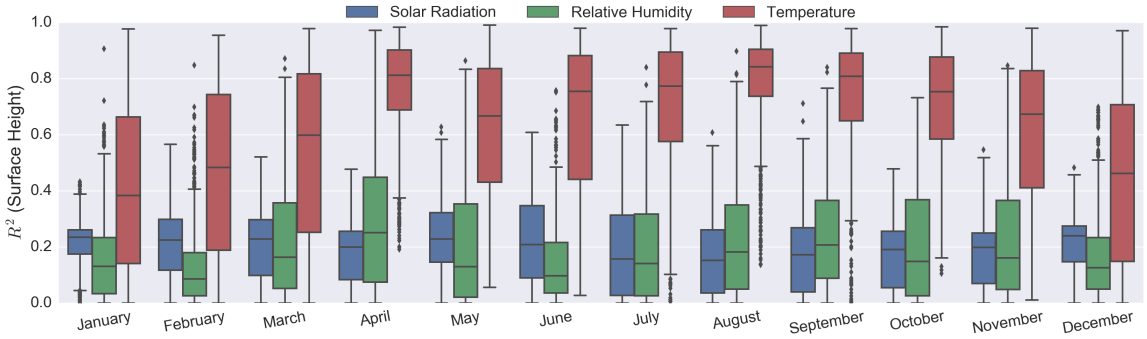
(a) Temperature (K) (b) Relative Humidity ($\%$)



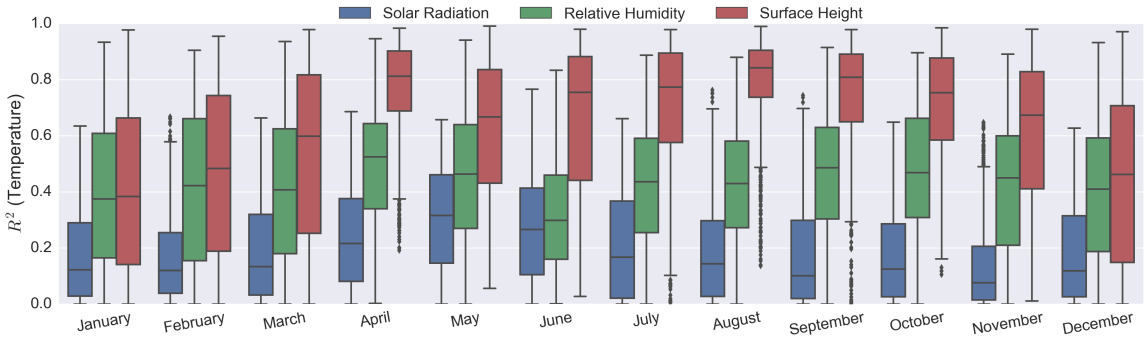
(c) Solar Radiation (W/m^2)

Fig. 3.4 Example of three environmental parameters that are utilised in this work: (a) temperature; (b) relative humidity; and (c) solar radiation. The colour bar on the right of each image corresponds to each parameter's value and unit.

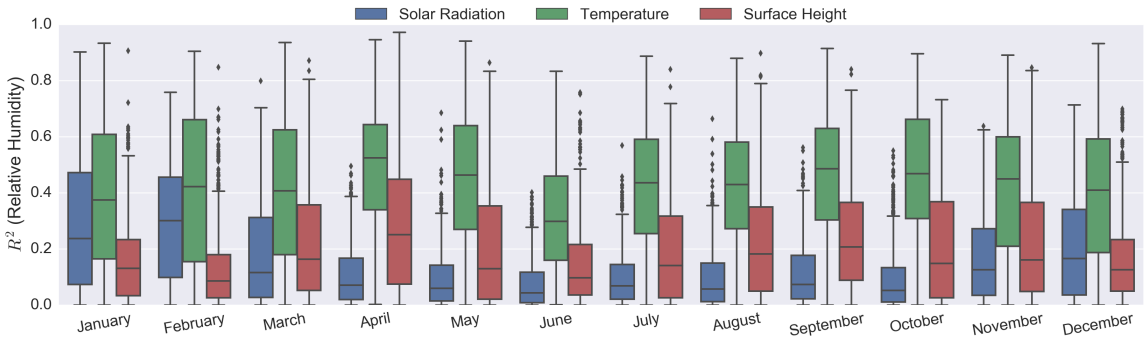
other parameters where the median of monthly R^2 value throughout the year lies below 0.4 Figure 3.5d.



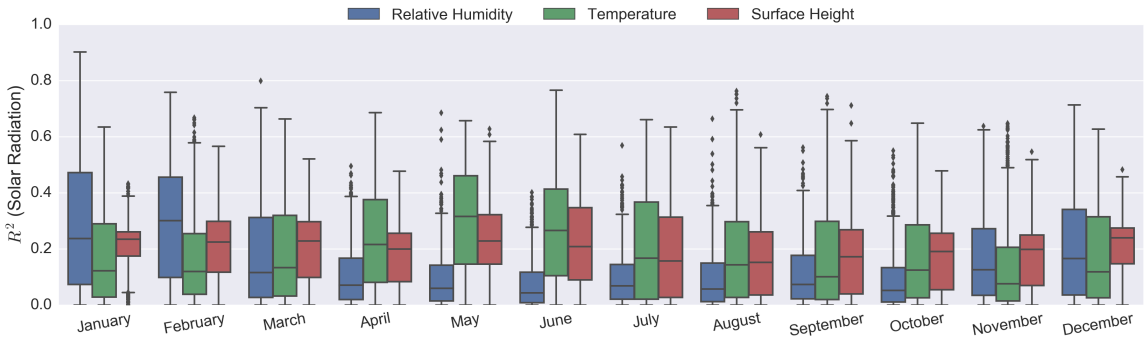
(a) Surface Height



(b) Temperature



(c) Relative Humidity



(d) Solar Radiation

Fig. 3.5 Box plot demonstrating the monthly R^2 value of hourly data (2-D map) between parameters throughout the year 2013. The horizontal axis represents the months in a year and the vertical axis shows the corresponding R^2 value.

3.1.2 Bee Experimental Data Set

CSIRO's Swarm Sensing project conducted an initial bee experiment at Geeveston, Tasmania, from 02 April to 27 November 2014. A radio-frequency identification (RFID) reader and a mini computer were installed at each hive entrance so that the presence of tagged bees (going in or out of the hive) near the reader will be detected and recorded in a Comma Separated Values (CSV) format output file. The data consists of two columns: (i) *datetime* data, that records the exact date and time when an individual is detected by the RFID reader; and (ii) *bee_id*, the unique identifier for an individual bee.

Four bee hives were set up and the Swarm Sensing team members visited the field twice a week (on average) to deploy the RFID tags each bee's thorax and, eventually, thousands of bees were tagged to collect bee detections data throughout the experiment.

3.1.3 Bee Foraging Flight Model

The data set in this section is mainly obtained from the generated output from [80]. The data set is only used for the artificial bee's foraging flight paths within the RoI.

Within the flight model, individual bees fly out from the hive along various random walk flight paths, with the characteristics or each flight path determined by the foraging role of that bee, the foraging activity that it is undertaking, the location of the bee within the foraging range and the environmental conditions currently being experienced by the bee.

The flight behaviour of the bees is determined by a set of rules, coded into the Insect Flight Simulator, which specifies the action to be taken by each bee, at each simulated time step, depending on its current foraging activity and the local environmental conditions. Other rules determine when the bee will change its foraging activity, depending on its location and environmental conditions.

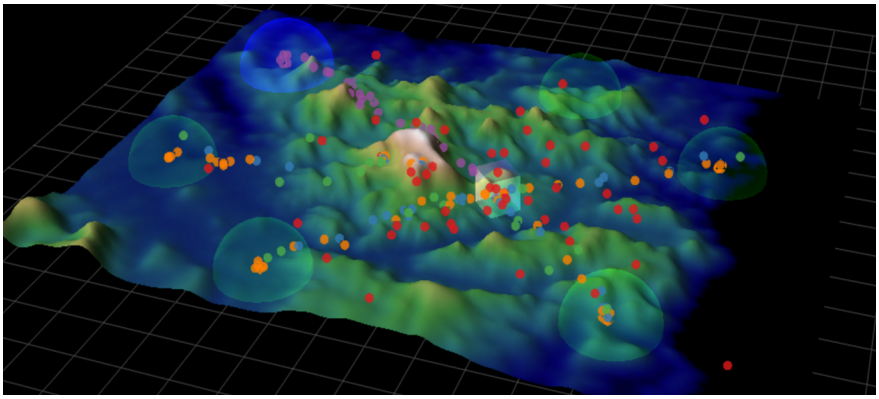


Fig. 3.6 A snapshot visualisation of the output generated by the agent-based computational bee foraging flight paths.

The simulation produces, as output, a set of realistic three-dimensional honey bee forager flight paths, embedded within a three-dimensional sensing environment. These flight paths can be visualised using three-dimensional animation software or used, as input, for various third-party visualisation and analysis software packages. An example is shown in Figure 3.6.

3.1.4 Benchmark Data Set

Finally, the weather stations data within the RoI (i.e., the South Esk's region, Tasmania) was utilised as the 'benchmark' data set. The data set is the initial sources for the 'modelled' high-resolution data set (Section 3.1.1). Table 3.4 shows the list of sites (weather stations) and their corresponding configurations that were employed for the purpose of this work:

Site#	Site name	Coordinates (lon, lat)	Parameters	Frequency
1	Ben Lomond	147.6613, -41.5401	temp	15m
2	Ben Ridge Road	147.7088, -41.3519	temp, rh	10m
3	Snow Hill Farm	147.8374, -41.8559	temp, rh	5m
4	St. Patricks Head	148.2178, -41.5770	temp, rh	10m
5	Storys Creek	147.7393, -41.6346	temp, rh	10m
6	Tom Gibson Nature Reserve	147.3031, -41.7729	temp, rh, rnet	15m

Table 3.4 Sites information that was utilised as the 'benchmark' data set. Note that within the 'Parameters' column, the following abbreviations are used: temp (temperature), rh (relative humidity), and rnet (solar radiation).

The main purpose of utilising this data set was to validate the spatio-temporal interpolation technique proposed in this dissertation. The swarm sensing simulation will be conducted by employing the 'modelled' environmental data. It would also be interesting to investigate the performance of the interpolation technique by comparing results with the 'benchmark' data set which was originally used for the creation of 'modelled' data. Such efforts are feasible because the proposed spatio-temporal interpolation algorithm in this work is capable of estimating values at any irregular space and time point.

3.2 Optimisation Algorithm - ESN Deployments

The deployments of Environmental Sensor Networks (ESN) is an interesting research area in which a particular algorithm (*spatial sampling* method) is used to obtain optimal placements of nodes within the RoI. As already mentioned in the previous chapter (Chapter 2, Section 2.1), a number of techniques have been investigated to address this problem, and yet, no optimal method is applicable and appropriate in all circumstances. Therefore, a new spatial-sampling technique is utilised based on Evolutionary Algorithm (EA) in this dissertation to build a near-optimal ESN [10]. EA mimics the procedure of the biological mechanism such as reproduce, cross over, mutation, recombination, and elitism, that maximise or minimise a user-defined *fitness function* in order to reach a near-optimal solution.

3.2.1 Problem Statement

Distinct types of static sensor nodes are denoted as: $SN_i = \{sn_{i,1}, sn_{i,2}, \cdots, sn_{i,n}, \cdots, sn_{i,N}\}$ with $i = item = \{hive, food, water, corner\}$ to represent bee hive, food source, water source, and weather stations located at the map's corners (top-left, top-right, bottom-left, and bottom-right) respectively. We can also denote nodes within entire networks using $SN = \{sn : sn \in SN_{hive} \cup SN_{food} \cup SN_{water} \cup SN_{corner}\}$.

Furthermore, for the purpose of this work, a constraint is being introduced for the optimisation algorithm so that the sensor nodes placements to be optimised (SN_{hive}) are located within the 'convex hull' generated by $SN_{food} \cup SN_{water}$ (Figure 3.7). This assumption is made in order to concentrate the bee hives in the middle-part of the RoI and to reduce the probability of insects flying around the map's border. Therefore, a solution will be invalid if one or more optimised sensor nodes (hives) is located outside of the convex hull (a grey area in Figure 3.7).

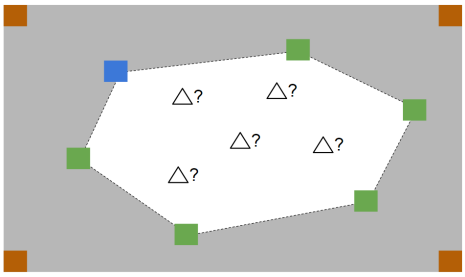


Fig. 3.7 Illustration of the 'convex hull' generated by $SN_{food} \cup SN_{water}$ (white area), in which the SN_{hive} is allowed to be optimised within. The squares are: weather stations at the map's corners (brown), food (green) and water (blue) sources respectively; whilst the triangle denotes the locations of bee hives to be optimised.

3.2.2 Chromosome Design

Chromosome (also called an 'individual' in EA) design is an important step in the development of any optimisation EA technique in order to represent a single solution so that it is assessable for quality evaluation using the so-called *fitness* function.

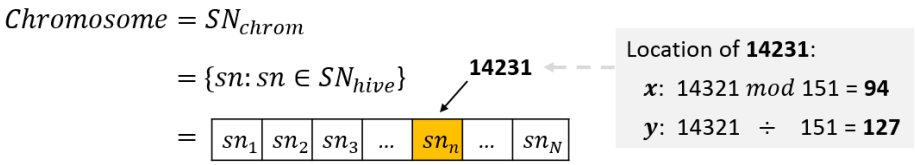


Fig. 3.8 A chromosome encoding and decoding example that consists of only the hive nodes which are to be optimised.

Let $SN_{chrom} = \{sn : sn \in SN_{hive}\}$ be a set of sensor node locations encoded as a chromosome for optimisation purposes. It stores a list of integers with the $N = N(SN_{chrom})$ elements, where each item ranges from 0 to 15250 (a total of $101 \times 151 = 15251$ grids/cells throughout the RoI). In order to obtain the exact spatial location (x and y) within the landscape (chromosome decoding), we use the following:

$$loc(c_{sn}) = \begin{cases} x = sn_n \bmod 151 \\ y = sn_n \div 151 \end{cases} \tag{3.1}$$

An example is shown in Figure 3.8 to illustrate the procedure. Then, a *fitness* function is used for quality assessment of a single chromosome in the following sub-section.

3.2.3 Fitness Function

The main motivation of the fitness function calculation is based on the fact that spatial interpolation is often criticised for an inability to estimate extreme values [84]. For example, within a mountainous region, if we only deploy the sensor nodes within the lower ground of the landscape (the foothills in this case), no interpolation method is capable of estimating the temperature values on the mountain's peak. Thus, the main objectives of the fitness function are focusing on the following:

- To capture the *representative* nodes within the RoI, such that extreme low and high values are represented. Here, 'representativeness' is defined as the ability of an interpolator to best estimate the condition of a particular environmental parameter of the RoI at a later stage (given a pre-defined number of nodes for optimisation).

- To minimise *redundant* nodes. The term 'redundant' in this case refers to those points that are able to be estimated by the interpolation method with less statistical error (e.g. RMSE);
- Create a *sparsely* distributed network of sensor nodes. In addition to the objectives given above, a dispersed sensor network design is also preferred.

A statistical error measurement method is exploited; Leave-One-Out Cross-Validation (LOCCV), in conjunction with spatial interpolation method to obtain the critical points over the landscape. The 'critical' points are defined as the representative sensor nodes placed in such a way that each node's observation is unable to be restored (estimated) using a conventional interpolation technique; In other words, the absence of any node would degrade the representativeness of the entire sensor network:

$$LOCCV(\hat{f}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_{chrom,n} - \hat{f}^{(-n)}(x_{chrom,n}))^2} \tag{3.2}$$

where $\hat{f}$ is a particular interpolation technique; N is the total number of nodes within SN_{chrom}; y_n is the observed value at $SN_{chrom,n}$; and $\hat{f}^{(-n)}(x_{chrom,n})$ is the estimated value of $x_{chrom,n}$ using $\hat{f}$ method using the sensor nodes of food sources, water source, and the hives (nodes to be optimised) with n^{th} node absent (i.e. $(SN_{food} \cup SN_{water} \cup SN_{hive}) \backslash SN_{hive,n}$). Also, note that the $LOOCV$ is only calculated from SN_{chrom}.

The $fitness$ function within the optimisation process is calculated by maximising the following equation:

$$\begin{aligned} fitness &= LOCCV(\hat{f}) \times sparsity \\ &= LOOCV(\hat{f}) \times min\{\ pdist(SN_{food} \cup SN_{water} \cup SN_{chrom})\ \} \end{aligned} \tag{3.3}$$

where the network's sparseness is estimated by obtaining the minimal pairwise distance ($pdist$) between the sensor nodes: $SN_{food} \cup SN_{water} \cup SN_{chrom}$.

3.3 Bee Behaviour Modelling

This section proposes the data-driven artificial bee simulation framework using the 'experimental' data set (DataSet 2, Section 3.1.2). In summary, the main components to be addressed are: (a) The start and end of single bee activity within a day; (b) The possible activities of a bee depending on time-of-day; (c) Propose the data-driven bee modelling framework. Consequently, the procedure to simulate an artificial bee will be discussed at the end.

3.3.1 Bee Behaviour Classification

First of all, the data set was 'grouped' based on individual bee on a daily basis. Then, the data needed to be analysed so that we could obtain meaningful information for analytic purposes of the data set. Bee activities are classified into three different behaviours to distinguish the possible activity of an individual bee throughout the day, namely: By The Entry (BTE), Short Mission (SM), and Foraging (FG). The FG activity includes activities occurring outside or inside of the hive. Table 3.5 summarises the categorisation rules for this work based on field observations:

	Successive Readings Threshold	Activity Duration
By the entry	$x < 3m$	$x < 30m$
Short mission	$3m \leq x \leq 6m$	$3m \leq x \leq 6m$
Foraging	$x > 6m$	$6m \leq x \leq 6h$

Table 3.5 Summary of bee activity classification rules based on field observations. Within the 'Activity Duration' column, if the duration x is more than the pre-determined range (e.g., by the entry $x > 30m$), the data will be omitted. Note that the configurations may vary for different bee species.

By The Entry (BTE). A detections difference that is less than $180s$ is to be categorised as BTE in which the bee hovers around the RFID reader making consecutive readings. Therefore, the estimated duration for this particular activity is made up of the accumulated values for those readings. This method allows us to detect dead bees near the reader, indicated

by a very high reading frequency of consecutive readings for the entire day, enabling these data to be classified as invalid and omitted from the model.

Short Mission (SM). A bee is assumed to be in 'short mission' where the interval of successive readings is between $3m$ and $6m$. During this period, bees are likely to be doing the following activities: short flights, orientation flights to familiarise themselves with the hive's surroundings (occurs among novice bees), walking around the hive, etc.

Foraging (FG) period. Successive readings with more than six minutes interval ($x > 6m$) will be classified as foraging; nevertheless, this activity is further divided into two categories: (a) Out-of-hive foraging where bee is scouting/exploiting for food/water sources; and (b) In-hive foraging period where an individual deposits food sources into the hive. In addition, it is also assumed that an individual bee does not forage before sun rise or after sun set.

Eventually, by applying the above behavioural rules to the 'experimental' data, we can obtain knowledge of the bee activities throughout the day and this information will be used in the following sub-sections.

3.3.2 Bee Activities Modelling

This sub-section is one of the main contributions of this dissertation that involves utilising the activity data generated from the previous sub-section (Section 3.3.1) to perform the curve fitting optimisation procedure to model bee behavioural activity in a day. To begin with, problem quantification (mathematical notation) needs to be formalised and will be used within the rest of the framework description. The data set (D) will be divided into two hierarchical levels (summarised in Table 3.6):

i *Bees' activities.* Let $A = \{a_{i=1}, a_{i=2}, a_{i=3}\} = \{BTE, SM, FG\}$ be a set of bee activities with a total of 3 items ($I = 3$) as described in Section 3.3.1.

ii *Time of day.* The second hierarchical level is the time-in-day $T = \{t_1, t_2, \cdots, t_j, \cdots, t_J\}$ which holds pre-defined 'bins' throughout the day to indicate occurrence (frequency) of a particular activity at a designated period of time. For the purpose of this work the bin size $t_{bin} = 30m$ is used which means that J (the number of elements in T) $= 24h \div 30m = 48$ elements (i.e. $T = \{t_1, t_2, \cdots, t_j, \cdots, t_{J=48}\}$).

A set of mathematical notations is also established in order to select a sub-set of data D based on a particular category (hierarchical level) of D. A number of examples are presented

Set Notation	Description	Representation		
		Item	Index	Example
A	Bee Activity	a_i	$i,\cdots,I$	$A=\{a_1,a_2,\cdots,a_i,\cdots,a_I\}$
T	Time of day	t_j	$j,\cdots,J$	$T=\{t_1,t_2,\cdots,t_j,\cdots,t_J\}$
D	Data set	$d_{i,j}$		$D=\{\cdots,d_{i,j},\cdots,d_{I,J}\}$

Table 3.6 Mathematical notations that will be used throughout this sub-section.

as follows:

$$\begin{aligned}
D &= \{d : d \in D\} \\
&= \{d_{1,1}, d_{1,2}, \cdots, d_{i,j}, \cdots, d_{I,J}\} \\
D_i &= D_{i=x} \\
&= \{d : d \in D \wedge i = x\} \\
&= \{d_{x,1}, d_{x,2}, \cdots, d_{x,j}, \cdots, d_{x,J}\} \\
D_j &= D_{j=x} \\
&= \{d : d \in D \wedge j = x\} \\
&= \{d_{1,x}, d_{2,x}, \cdots, d_{i,x}, \cdots, d_{I,x}\}
\end{aligned}$$

Also, please note that x is an artificial notation which serves different purposes. For example, in case of $D_i = D_{i=x}$, the notation x represents a_i in activity A.

A summation function $S(d) = \sum_n^N d_n$ over a set of datum d that consists of N elements. Further, the weighted mean (μ^*) and standard deviation (σ^*) are also formalised to correspond with the data availability ($d_{i,j}$, here, denoted by w_i):

$$\mu^*(V,W) = \frac{\sum_i^N w_i \cdot v_i}{\sum_i^N w_i} \tag{3.4}$$

$$\sigma^*(V,W) = \left[\frac{\sum_i^N w_i \cdot (v_i - \mu^*)^2}{\sum_i^N w_i}\right]^{\frac{1}{2}} \tag{3.5}$$

where V and W are the lists of values (v_i) and weights (w_i) respectively, with a total of N elements.

Modelling Bee Activity in Time-of-day

By employing the data set D previously discussed, we model the bee activity within a day using the following Gaussian Probability Density Function (PDF):

$$G_{ALL} = \sum_{i}^{I} G_i \quad = G_{BTE} + G_{SM} + G_{FG} \tag{3.6}$$

$$G_i(x, I, T_\mu, T_\sigma) = Ie^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.7}$$

where G_i is one single Gaussian distribution function of the i^{th} activity in A, x is the data point (time of day in this case) to be estimated, I is the intensity, and μ and σ are the mean and the standard deviation of the distribution respectively. Then, the curve fitting optimisation process will be executed using the Gaussian PDF to obtain the activity distribution within a day.

To do this, the preliminary step is to 'normalise' the data because a particular bee activity (especially BTE) could occur any moment regardless time of day and will cause the so-called 'background effect' (BKG) of the data which will result in degradation of the curve fitting performance. The presence of such an effect does not comply with the characteristic of a PDF and must be removed from the data before executing the optimisation. BKG can be estimated by calculating the mean (μ) of D_i that holds the minimal Coefficient of Variation (CV) of the first n_{first} and last n_{last} datum within a day:

$$\underset{n_{first}, n_{last}}{\arg\min}\ CV\big(\ (D_i)_{n_{first}} \cup (D_i)_{n_{last}}\ \big) \tag{3.8}$$

where $CV()$ is the CV function; the n_{first} datum $(D_i)_{n_{first}} = \big\{S(D_{i,j=z}) : z \in \{1, 2, \cdots, n_{first}\}\big\}$; and the n_{last} datum $(D_i)_{n_{last}} = \big\{S(D_{i,j=z}) : z \in \{L - n_{last}, \cdots, L\}\big\}$. Then, the BKG can be estimated by:

$$BKG_i = \mu\big(\ (D_i)_{n_{first}} \cup (D_i)_{n_{last}}\ \big) \tag{3.9}$$

with n_{first} and n_{last} corresponding to the minimised Equation 3.8. This framework also assumes that an individual bee only forages during the day (i.e. in between sunrise and sunset); Therefore, the BKG for foraging period (FG) does not exist. Then, the Gaussian distribution to be curve-fitted (Equation 3.7) can now be re-written as:

$$G_i(x, BKG, I, T_\mu, T_\sigma) = Ie^{-\frac{(x-T_\mu)^2}{2T_\sigma^2}} + BKG \tag{3.10}$$

Then, let time $T_{first,last} = \{t_x \in \mathbb{Z} \wedge n_{first} < x < n_{last}\}$ be a sub-set of T and its corresponding datum $U_i = \{D_{i,j=x} : x \in \mathbb{Z} \wedge n_{first} < x < n_{last}\}$. The remaining parameters for the individual Gaussian G_i of distinct activity (a_i) are estimated in the following:

	Estimation	Constraint
I_i	$\mu(\{u : u \in U_i \wedge u \geq U_{i,unq,(n-2)}\})$	$\sigma(\{u : u \in U_i \wedge u \geq U_{i,unq,(n-2)}\})$
$(T_\mu)_i$	$\mu^*(\{T_{first,last}, U_i\})$	30 *minutes*
$(T_\sigma)_i$	$\sigma^*(\{T_{first,last}, U_i\})$	$\sigma^*(\{T_{first,last}, U_i\}) \div 2$

where u denotes each datum within U_i and $U_{i,unq,(n-2)}$ is the third largest 'unique' value within U_i (represented using an order statistic). Lastly, the lower and higher boundaries (i.e. search spaces) for the optimisation are calculated by:

$$\begin{aligned} C_{lo} &= Estimation - Constraint \\ C_{hi} &= Estimation + Constraint \end{aligned} \tag{3.11}$$

Chromosome Design

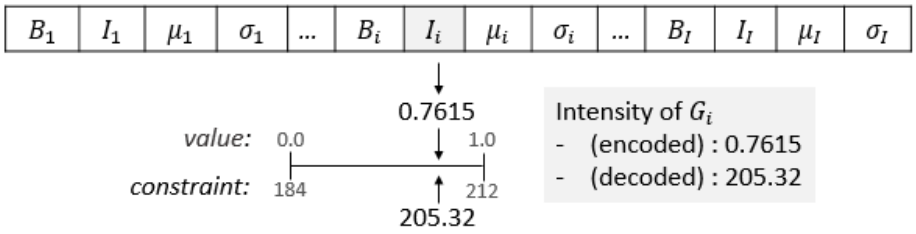


Fig. 3.9 An illustration of the chromosome design utilised in this work, where each element within the individual holds a value between 0 and 1, and B denotes the background (BKG). The encoding and decoding example of the intensity (I_i) parameter of distribution G_i is also presented. In this case, assume that we have element I_i with value 0.7615 (encoded) within the individual which is equivalent to 205.32 (decoded) after applying Equation 3.12.

The process of chromosome encoding and decoding quantifies the problem creating an 'individual' for the optimisation process. For the purpose of this curve fitting, a single individual is designed using the a similar approach as to that [85]. A single parameter to be optimised will be encoded as a value between 0 and 1, the following equation is utilised to decode the value:

$$p(x, C_{lo}, C_{hi}) = C_{lo} + x \cdot (C_{hi} - C_{lo}) \tag{3.12}$$

where x and p are the encoded and the decoded value of a particular Gaussian parameter; C_{lo} and C_{hi} are the constraint values (lower bound and higher bound) calculated from Equation

3.11. Figure 3.9 demonstrates the design of single individual with a decoding example being a parameter I of the Gaussian distribution G_i.

Fitness function

In this work, the optimisation will minimise the sum of chi-square (χ^2) fitness function to obtain the best individual [10, 85]:

$$fitness = \sum_{i}^{I} \chi_i^2 \tag{3.13}$$

$$\chi_i^2 = \frac{1}{J - N_p} \sum_{j}^{J} \frac{(d_{i,j} - G_{i,j})^2}{d_{i,j} + 1} \tag{3.14}$$

where I and J are the total number of activities A and time of day T respectively, N_p is the number of parameters to be optimised (4 in this case) for each Gaussian, $d_{i,j}$ and $G_{i,j}$ are the data and curve-fitted lines at i^{th} activity and j^{th} time-of-day. Note that the $+1$ within the denominator of Equation 3.14 is introduced to avoid a divide-by-zero error.

3.3.3 Artificial Bee Simulation

This section will describe the procedure to simulate a bee within the field by utilising the previously discussed components based on Figure 3.10.

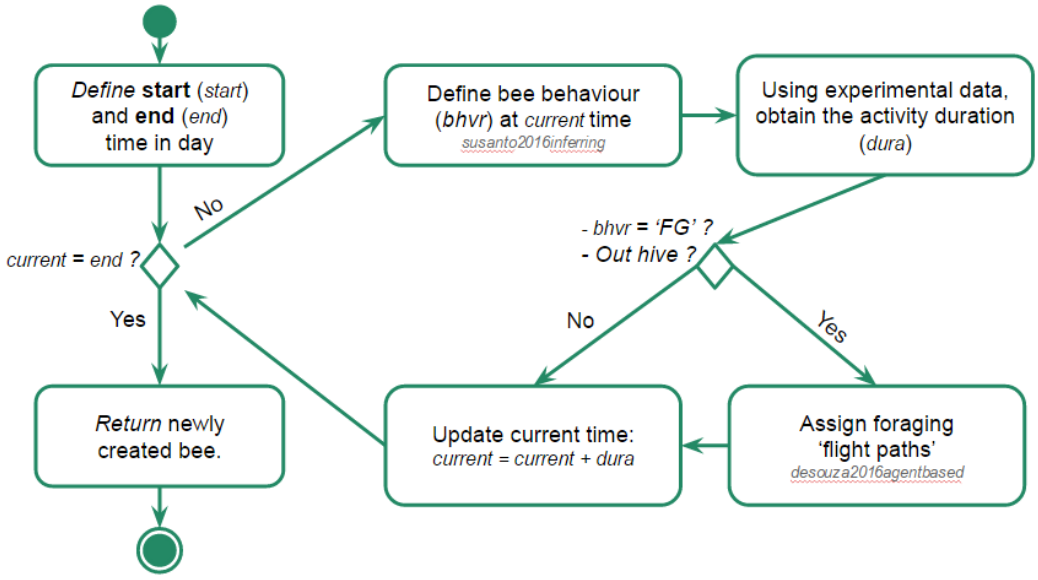


Fig. 3.10 Simulation procedure to generate an artificial bee.

The initial step is to obtain the *first* and *last* detections ($detection_{first}$ and $detection_{last}$) of the bee within a day. To do this, two time-stamps will be randomly chosen using distribution G_{ALL} (Equation 3.6) and the lower/higher value will become the first/last detection for that particular bee.

In the following, operator t_{curr} is used to indicate the current time-stamp of the bee to be simulated within the iterations. Provided $t_{curr} \neq detection_{last}$, the algorithm obtains a list of 'possible' activities (Section 3.3.2). Here, 'possible activity' means that the *next* bee behaviour is based on the empirical rule illustrated in Table 3.7. Then, by utilising the optimised (curve fitted) PDF of distinct 'possible' activities, a 'weight' is assigned according to the current time-of-day t_{curr}. Eventually, the bee's activity at t_{curr} is randomly chosen from the weighted choices of activities. In addition, in case where $a_{next} = FG$ and where the bee is out-of-hive, a bee foraging flight paths (Section 3.1.3) will also be allocated to the particular bee activity. The random activity *duration* is given based on the information in the post-processed 'experimental' data set (Section 3.3.1). The entire iteration is repeated if $t_{curr} \neq detection_{last}$; otherwise, the simulation process returns the newly generated artificial bee.

a_{curr}	Activity Description	Possible next activity (a_{next})
$Init$	Initial/first detection	$a_{next} = \{BTE, SM, FG\}$
BTE	By The Entry	$a_{next} = \{SM, FG\}$
SM	Short Mission	$a_{next} = \{BTE, FG\}$
FG	Foraging	$a_{next} = \{BTE, SM\}$

Table 3.7 The empirical rules utilised for choosing the next activity (a_{next}) of a bee based on its current activity (a_{curr}). The *Init* denotes that the t_{curr} is the first detection ($detection_{first}$) of that particular bee on that day.

3.4 Swarm Sensing Data Sampling

The entire swarm sensing field simulation (hybrid ESN) can be generalised in the following:

Step 1 Initialise a number of pre-defined static nodes (acting as weather stations) within the region of interest, they are: food (SN_{food}) and water sources (SN_{water}), and sensor nodes at the map's corners (SN_{corner}).

Step 2 Execute the optimisation algorithm for ESN deployments (Section 3.2) to locate the *statistically optimised* locations for a number of pre-defined hives (SN_{hive}) within the area under study. Please note that, for the purpose of this work, the optimisation execution is also being constrained in a way that the locations of individual hive in SN_{hive} are to be located within the convex hull created by $SN_{food,water} = SN_{food} \cup SN_{water}$.

Step 3 Employ the artificial bee simulation framework as described in Section 3.3 from each hive (SN_{hive}) in order to generate a user-defined number of bees. The bees are acting as mobile sensor nodes that collect high frequency data as they move (forage) throughout the landscape.

Hence, by utilising the South Esk Modelled data set (Section 3.1.1) the main objective of this section is to propose a set of procedures to perform data sampling for the field simulation described above which corresponds to different sensor node types (i.e. static and mobile) as illustrated in Table 3.2.

3.4.1 Fixed Sensor Nodes

Since the field simulation is using the South Esk 'modelled' data set (Table 3.3) that consists of hourly data within the RoI, the simulated weather stations $\left(SN_{static} = SN_{food} \cup SN_{water} \cup SN_{hive}\right)$ sense the environment in a way that collect environmental parameter data on an hourly basis. Note that the fixed sensor node deployments (spatial sampling) will utilise the optimisation approach described in Section 3.2.

3.4.2 Mobile Sensor Nodes

On the other hand, for the purpose of this dissertation, the mobile sensor will be mounted on the insect's thorax and collects environmental data in a way that follows empirical rules according to the insect's behaviours (Table 3.8)

	Bee Behaviour			
	By The Entry	Short Mission	Foraging (out of hive)	Foraging (in hive)
Bee's Location	Hive entrance	Around hive	Out of hive	Inside hive
Detections	start $\cdots$ end	start, end	start $\cdots$ end	start, end
Frequency	$10s$	N/A	$60s$	N/A
Activity	Hive defence	Orientation flights	Search / exploit food sources	Deposit pollen

Table 3.8 Data sampling configuration for mobile sensor nodes – insects. Note that the '$\cdots$' on row 'Detections' indicates that the data are collected continuously with a certain interval (column 'Frequency') between the *start* and *end* time-in-day.

First, if the bee is detected by-the-entry (BTE) of the hive, a very high frequency of readings will be carried out (i.e. 10 *sec* interval). This is based on the assumption that a reader will be installed at the hive's entrance, in which has a similar configuration to that used in the CSIRO's Swarm Sensing initial bee experiment (Section 3.1.2).

In the case of a Short Mission (SM), which is believed to occur only among novice bees learning to fly, only the *start* and *end* time-of-detection at the hive's entrance will be recorded.

Lastly, the foraging (FG) behaviour represent the actual bee 'foraging' periods. Such a behaviour includes *in-hive* FG and *out-of-hive* FG. An *out-of-hive* FG is where a consecutive 'sensing' of the environment will take place as bees move throughout the landscape. A bee flight path will also be assigned to such a behaviour using the bee foraging flight model data set (Section 3.1.3). In this simulation, $\approx 60sec$ interval of detections is utilised because by considering the hardware capability limitations (i.e. power source), it is more realistic that sensors further away from the reader will have lower reading frequency compared with nearer sensors to the readers, for instance, BTE. On the other hand, similar to SM, the *in-hive* FG will only record data at the start and end of the behaviour period. In this case, the bees are assumed to be involved returning from an *out-of-hive* FG to deposit food or water and performing the so-called 'waggle dance' to inform other bees of the newly found food/water sources.

By having the above bee *out-of-hive* foraging configurations, the simulated bee's spatial and temporal 'sensed' data will be irregularly spaced. However, since the South Esk Modelled

data set (Section 3.1.1) consists of only regularly spatial (location represented as integers) and temporal data (on an hourly basis), further estimation of irregularly-spaced environmental data is necessary from the data set.

In this work, a linear reduction-based spatio-temporal interpolation algorithm is employed [45]. The term 'reduction' implies that the algorithm first interpolates the value in the spatial dimension, then 'reduces' it to an estimation in the temporal dimension.

Algorithm 1 Estimate the value in the spatial dimension

function SPATIAL(x, y, t)
  if x **is not** integer **and** y **is not** integer **then**
    $K \leftarrow \{(x_{lo}, y_{lo}, t), (x_{lo}, y_{hi}, t), (x_{hi}, y_{lo}, t), (x_{hi}, y_{hi}, t)\}$
  else if x **is** integer **and** y **is not** integer **then**
    $K \leftarrow \{(x_{lo}, y_{lo}, t), (x_{lo}, y_{hi}, t)\}$
  else if x **is not** integer **and** y **is** integer **then**
    $K \leftarrow \{(x_{lo}, y_{lo}, t), (x_{hi}, y_{lo}, t)\}$
  else
    $K \leftarrow \{(x, y, t)\}$
  end if
  return $IDW_s(\ (x, y),\ K\ )$
end function

To do this, Algorithm 1 is employed in conjunction with the Inverse Distance Weighting (IDW) interpolation technique to compute the value at spatial dimension (x, y) and time t:

$$IDW_s\big((x,y),K\big) = \sum_{i}^{N} w_i \cdot v_i \quad , w_i = \frac{1}{d_i^a} \tag{3.15}$$

where K is the sample data that will be considered in the interpolation process; d_i, w_i and v_i are the distance, weight and value of the i^{th} sample point respectively; and a is the *distance-decay* parameter. Furthermore, based on Algorithm 1, dim_{lo} and dim_{hi} is the lower and upper integer value of a particular dimension $dim \in \{x, y\}$; In other words, this method only considers the nearest data in the linear spatial estimation. Then, the 'reduction' step to the irregular-temporal dimension t is performed using:

$$f_{st}(x,y,t) = spatial(x,y,t_{lo}) \times \frac{t_{hi} - t}{t_{hi} - t_{lo}} + spatial(x,y,t_{hi}) \times \frac{t - t_{lo}}{t_{hi} - t_{lo}} \tag{3.16}$$

where $spatial()$ represents the Algorithm 1; t_{lo} and t_{hi} are the rounded-down and rounded-up values of time t. Note that if t is an integer, Equation 3.16 will simply become a spatial-only estimation, i.e. $f_{st}(x,y,t) = spatial(x,y,t)$.

3.5 Spatio-temporal Interpolation (STI) Algorithm

In this dissertation, a computationally efficient spatio-temporal interpolation (STI) technique is proposed to perform environmental modelling from a swarm of static and mobile sensor nodes produced by the swarm sensing field simulation (Section 3.4). Firstly, a new variation of the method is proposed to model the space-time interaction by utilising a geo-statistical technique for the weighting mechanism of the STI algorithm to be discussed shortly.

Then, the high-resolution irregularly-spaced observations generated from the RoI will be pre-processed for the STI technique. The rationale of this procedure is to 'reduce' the complexity of sample points to be processed by the STI algorithm to facilitate computational efficiency. On the other hand, such activity can also be seen as Quality Control (QC) because the STI algorithm will consider the data quality in terms of data *variability* and observation *density*.

Lastly, a hybrid approach (using an extension and a reduction method) STI interpolation technique will be proposed that satisfies the swarm sensing requirements: (a) computationally not intensive; (b) acceptable quality; and (c) complies with the environmental situation in reality, such as, producing smooth/gradual space-time transitions. The error measurement of the corresponding STI algorithm will be given at the end. The following sub-sections will describe the aforementioned components.

3.5.1 Geo-statistical Modelling: Spatio-temporal Variogram

For this part, only the static nodes $SN_{static} = \{SN_{food} \cup SN_{water} \cup SN_{hive}\}$ will be utilised. In geostatistic, an observation can be written as:

$$Z(s,t) = m(s,t) + Y(s,t) \tag{3.17}$$

where $Z(s,t)$, $m(s,t)$, and $Y(s,t)$ are the spatio-temporal observation, mean component, and stochastic residual component respectively. In order to model the spatio-temporal variogram, only the $Y(s,t)$ component is incorporated to indicate the relationship between a particular space(h_s) and/or time (h_t) distance. The variogram is created by using:

$$\gamma_{st}(h_s,h_t) = \frac{1}{2}\left[Var\big(Y(s+h_s,t+h_t) - Y(s,t)\big)\right] \tag{3.18}$$

where $Var()$ is a function to calculate the variance of two observations with a specified spatio-temporal 'lag' distance (h_s and h_t). This work will utilise the generalised product-sum

model proposed by De Iaco [53]:

$$\gamma_{st}(h_s,h_t) = \gamma_{st}(h_s,0) + \gamma_{st}(0,h_t) + k\gamma_{st}(h_s,0)\gamma_{st}(0,h_t) \tag{3.19}$$

with

$$k = \frac{sill\gamma_{st}(h_s,0) + sill\gamma_{st}(0,h_t) - sill\gamma_{st}(h_s,h_t)}{sill\gamma_{st}(h_s,0)sill\gamma_{st}(0,h_t)} \tag{3.20}$$

where the *sill* at different dimensions (space, time, and spatio-temporal) are estimated by using the variance of the corresponding data set. Finally, the following rule must be met in order to satisfy the admissibility for γ_{st} as in Equation 3.19.

$$0 < k \leq 1/\max\{sill\gamma_{st}(h_s,0), sill\gamma_{st}(0,h_t)\} \tag{3.21}$$

3.5.2 Raw Data Pre-processing

Figure 3.11 illustrates the processing example for a 2D spatial-only case. The entire RoI will be divided into a number of partitions in a way that each partition consists of one sample data point on average. Then, each partitions holds the mean (μ) and standard deviation (σ) calculated from the sample data within the corresponding partitions and is spatially-located at the midpoint of rectangle (light-green as in the right of Figure 3.11). On other words, the interpolation algorithm will use the μ value for the estimation, while the σ is used to compute the estimation error (Section 3.5.4).

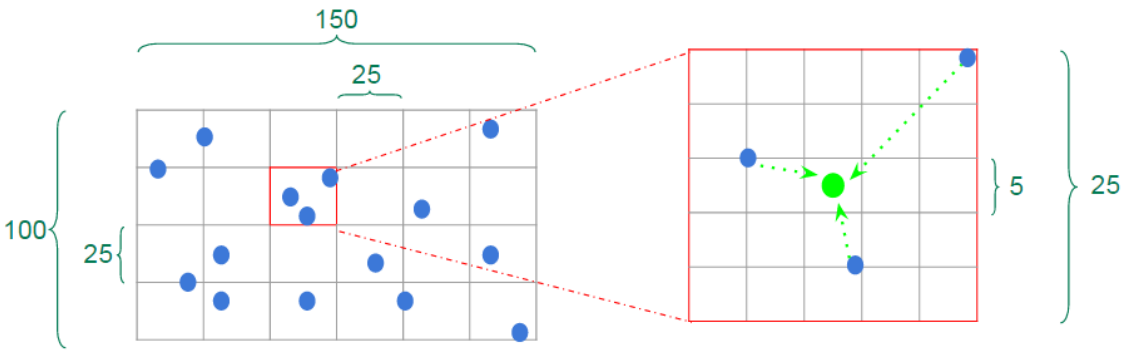


Fig. 3.11 Data pre-processing (spatial-only case). Left: 2D spatial map; Right: 'partitioned' sub-area from the map.

Such a procedure can be extended to the spatio-temporal case (2D-spatial and 1D-temporal) as demonstrated in Figure 3.12, where the actual swarm sensing field simulation high-resolution sample observations will be generated (left of the figure). Similar to the estimation presented on the right of Figure 3.11, on the right of Figure 3.12 is an example of

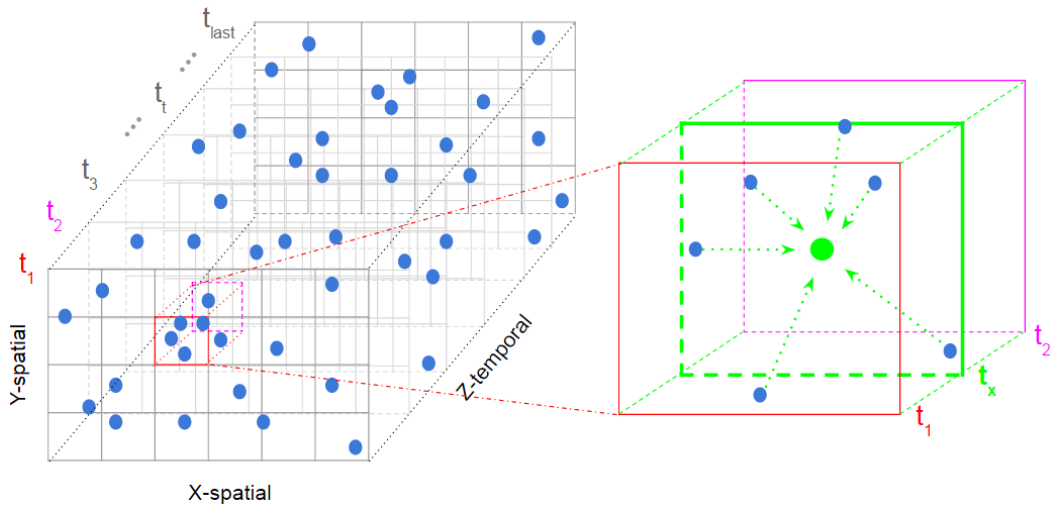


Fig. 3.12 Data pre-processing for 2D-spatial + 1D-temporal case. Left: 3D spatio-temporal data 'cube'; Right: 'partitioned' sub-area of the 3D data cube.

the partition value located at the midpoint of time $t_x = (t_1 + t_2)/2$ that holds the μ and σ of the observations (blue dots).

After execution, the raw data from the field simulation will be in the form of a regularly-spaced spatio-temporal grid (called the 'processed' data/observations hereafter) that holds the μ and σ values for the purpose of the STI estimation (Section 3.5.3) and its corresponding error measurements (Section 3.5.4) respectively.

3.5.3 The Hybrid Approach STI Algorithm

Let $V = \{v_{st,1}, v_{st,2}, v_{st,3}, \cdots, v_{st,i}, \cdots, v_{st,N}\} = \{\mu_{st,1}, \mu_{st,2}, \mu_{st,3}, \cdots, \mu_{st,i}, \cdots, \mu_{st,N}\}$ be a list of space-time 'processed' observations with a total of N elements:

$$STI_{red}(x,y,t) = STI_{ext}(x,y,t_{lo}) \times \frac{t_{hi} - t}{t_{hi} - t_{lo}} + STI_{ext}(x,y,t_{hi}) \times \frac{t - t_{lo}}{t_{hi} - t_{lo}} \tag{3.22}$$

$$STI_{ext}(x,y,t) = \sum_{i=1}^{N} w_{st,i} \cdot v_{st,i} = \sum_{i=1}^{N} \gamma_{st,i}^{-2} \cdot v_{st,i} \tag{3.23}$$

where STI_{red} and STI_{ext} are the reduction and extension approaches of the STI algorithm respectively; x, y and t are the (x,y) spatial-location at time t to be interpolated; t_{lo} and t_{hi} are the lower and higher time indices based on time t; and finally, γ_{st} is the spatio-temporal variogram that models the space-time interaction for the weighting mechanism (Section 3.5.1). An example of the above algorithm is illustrated in Figure 3.13.

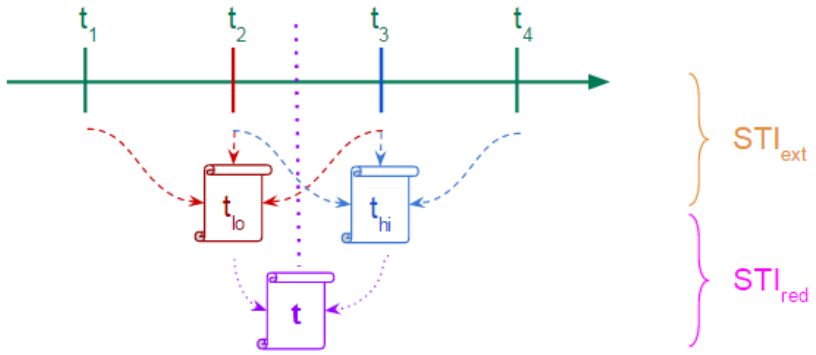


Fig. 3.13 Visual illustration of the proposed STI algorithm. In this example, the value at time t is to be interpolated. The algorithm first estimates the values at both $t_{lo} = t_2$ (red) and $t_{hi} = t_3$ (blue) utilising the extension approach (Equation 3.23); and then interpolates the value at time t (purple) using the reduction approach (3.22).

In this work, only plus-and-minus one time index (i.e. ± 1 hour in this case) will be utilised based on the justification that the observation at a particular time will not be correlated at the exactly same time-frame on the previous and/or following day. Furthermore, such a configuration has been shown to be superior to the empirical data set (South Esk Modelled data set in Section 3.1.1) which is the same data set as is used in this work [86]. Eventually, the spatio-temporal variogram (Section 3.5.1) is only modelled on a daily basis to facilitate computational efficiency.

3.5.4 STI Algorithm Error Measurements

Finally, the proposed STI algorithm is able to calculate the measurement error using the following equation (similar to Equation 3.23):

$$STI_{err}(x,y,t) = \sum_{i=1}^{N} w_{st,i} \cdot \sigma_{st,i} = \sum_{i=1}^{N} \gamma_{st,i}^{-2} \cdot \sigma_{st,i} \tag{3.24}$$

This calculation is mainly used to address the Quality Control (QC) component of the data utilised within the STI process, so that the proposed algorithm is able to provide a certain confidence level during interpretation.

3.5.5 Performance Assessment

There are many statistical techniques that can be used for performance evaluation of an interpolation method. Li and Heap [54] lists the mostly used statistical techniques for evaluation of interpolation methods. One of these techniques, the leave-one-out cross-

validation method will be implemented to evaluate the capability of interpolation method that has been developed based on: e.g., how well it can predict an omitted value, compared with the original value at that location [58, 72, 76].

The interpolation technique will also be assessed using Root Mean Square Error (RMSE). This is used to measure the error between the predicted values, compared with the model being estimated [49, 72, 75, 76, 87]. RMSE is based on the following equation:

$$RMSE = \left[\frac{1}{N} \sum_{i=1}^{N} (o_i - p_i)^2 \right]^{\frac{1}{2}} \tag{3.25}$$

where N is the total number of samples, o_i and p_i are the observed and the estimated/interpolated values of the i^{th} node. However, RMSE suffers from the drawback of being sensitive to outliers [54]. Therefore, the results could also be investigated using Mean Absolute Error (MAE), a measurement that is less sensitive to extreme values:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |o_i - p_i| \tag{3.26}$$

Li [54] suggests that in assessing an interpolation technique's performance, a combination of exact (cross-validation) and inexact (RMSE and MAE) methods is desirable to have confidence in the overall capability of the method.

The statistical mean will also be used to measure how 'correlated' two sets of variables are using the Pearson's r product-moment correlation of coefficient. The calculation is as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{3.27}$$

where $x = \{x_1, x_2, ..., x_n\}$ and $y = \{y_1, y_2, ..., y_n\}$ are the two sets of values being examined; $\bar{x}$ and $\bar{y}$ are the mean value of x and y respectively. The resultant value will always lie between -1 (total negatively correlated), 0 (no correlation) and +1 (perfect positive correlation). In this study, x and y can be seen as the observed (sample points) and estimated values respectively.

3.6 System Design

3.6.1 Software

Python programming language was used for this development work. This particular language was particularly chosen because it allows a user to build the research prototype and to obtain results in a relatively short period of time. Furthermore, because of an increase in the size of the Python community and the number of useful packages for scientific purposes (i.e. data processing and analysis, optimisation, visualisation, etc), it was considered that Python would be an ideal choice especially with time being a critical aspect for a PhD candidate (i.e. approximately 3 years).

The following Python packages were utilised during the software development process:

i *Numpy* [88]. A computationally effective mathematical computation package used in Python.

ii *Scipy* [89]. Provides a number of useful built-in algorithms for scientific computing.

iii *Pandas* [90]. A high performance, open-source Python data structure and analysis tools.

iv *Matplotlib* [91] and *seaborn* [92]. Widely used Python libraries for scientific plotting.

v *Distributed Evolutionary Algorithms in Python (DEAP)* [93]. An evolutionary computation framework used in Python for single or multi optimisation.

The entire field simulation and the spatio-temporal interpolation algorithm are implemented using the aforementioned tools. Eventually, the proposed framework in this dissertation is intended to assist environmental managers and/or researchers to make better decisions by using the software that has been developed – both statistical methods and visual analytics.

3.6.2 Hardware

The software development throughout the work was conducted on a personal computer with 3.40 GHz with 12GB RAM on a Windows 7 64-bit operating system.

Chapter 4

Software Implementation

This chapter discusses the design and execution of the experimental field simulation framework for the Swarm Sensing Project based on the previous chapter (Chapter 3: Methodology). The experiment was conducted using the following procedure:

Step 1 *Spatial sampling for near-optimal environmental sensor networks (ESN) deployments.* As previously discussed, the optimisation algorithm (Section 3.2) is mainly utilised for the deployments of hives' locations within the Region of Interest (RoI). Furthermore, the method also employed for the following purpose: comparing different spatial interpolation algorithms by considering several aspects, such that, statistical error, computational efficiency, and visual results acceptability.

Step 2 *Data-driven bee behavioural modelling.* This section covers the statistical modelling algorithm (curve fitting optimisation, Section 3.3.2) to generate artificial bees (Section 3.3.3) within the RoI that act as mobile sensor nodes flying through the landscape and 'sense' the environmental situation that they are experiencing.

Step 3 *Swarm sensing data sampling from the hybrid ESN field simulation.* Procedures to generate the high-resolution spatio-temporal environmental observations within the RoI from static (weather stations, hives, food and water sources) and mobile (e.g. insects) sensor nodes.

Step 4 *Evaluating the proposed spatio-temporal interpolation (STI) algorithms.* This section demonstrates the final results of high-resolution environmental modelling framework which has been proposed in this dissertation.

Each section consists of an introduction to various experimental designs and applications, followed by the corresponding simulation results and discussion.

4.1 Simulation 1: Spatial Sampling of Static Nodes

First of all, optimisation is executed mainly for the bee hives' location using the Evolutionary Algorithm (EA, see Section 3.2) to build a static-only ESN incorporating the pre-defined sensor nodes from the food sources, water source, and the map's corners (illustrated in Table 3.2).

However, as previously mentioned, this method will also be employed for the spatial sampling of different numbers of sensor nodes in order to evaluate the 2-Dimensional (2D) interpolation algorithms. Based on the assumption that we will collect a huge amount of data (up to thousands of observations) with a very high temporal frequency (every minute), computational efficiency is also a crucial factor during algorithm selection. Therefore, a careful analysis must be performed during selection to ensure a 'balanced' spatio-temporal interpolation algorithm is developed. The criteria for a 'balanced' spatial interpolation algorithm are as follows:

- *Computational efficiency*. The processing time does not increase significantly as the number of sensor nodes increases;
- *Produces acceptable statistical error*. This is the most widely compared element for spatial interpolation algorithm performance assessments;
- *Creates visually acceptable results*. Some interpolation techniques yield abrupt results, which do not faithfully represent real environmental situations.

4.1.1 Experimental Setup

This experiment includes two components: (i) ESN optimisation for the field simulation; and (b) ESN optimisation for comparing different interpolators.

Swarm Sensing Field Simulation. Using the description in Section 3.2, we simulate only five hives locations for this purpose, i.e. $N(SN_{hive}) = 5$. Also, SN_{hive} will be optimised within the 'convex hull' generated by static nodes: $SN_{food} \cup SN_{water}$.

Comparing different interpolators. The configuration for this experiment is different from that for the Swarm Sensing Field Simulation discussed above. In this case, the experiment does not incorporate the static nodes for food and water sources (i.e. $SN_{food} \cup SN_{water}$); it only uses the static nodes at the map's corners (SN_{corner}). Then, the ESN optimisation algorithm is executed based on different numbers of sensor nodes $N = \{5, 10, 15, 20, 25\}$. These

designs will be used to compare distinct spatial interpolators, namely: Ordinary Kriging (OK), Inverse Distance Weighting (IDW), and Shape Function (SF) (also called the Triangle Irregular Network – TIN). Finally, MAE (Equation 3.26) and RMSE (Equation 3.25) will be utilised for the statistical error validation.

The following table depicts the configuration employed for the EA:

Parameter	Value
Population size	50
Crossover probability	0.7
Mutation probability	0.05
Crossover operation	One point

Table 4.1 EA parameter configuration to be utilised within the execution.

4.1.2 Results

Bee Hives Locations for the Field Simulation

(a) $N(SN_{hive}) = 0$ (b) $N(SN_{hive}) = 5$

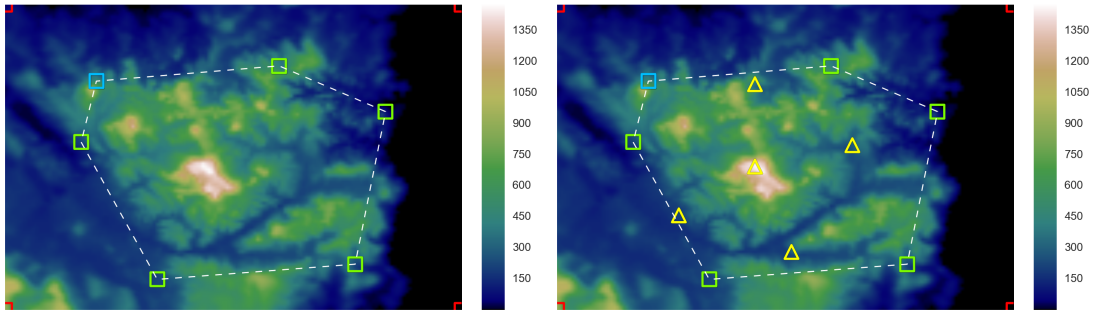


Fig. 4.1 Demonstration of the EA-assisted ESN optimisation: (a) the RoI and the predefined static nodes for execution; and (b) using the number of hives to be optimised $N = N(SN_{hive}) = 5$. The figure is labelled as follows: red square (SN_{corner}), green square (SN_{food}), blue square (SN_{water}), yellow triangle (SN_{hive}), and the 'convex hull' area (dashed-line connecting nodes: $SN_{food} \cup SN_{water}$) within which SN_{hive} are to be optimised..

The simulation has been executed and visualisation results for the optimised ESN deployment is presented in Figure 4.1. Figure 4.1 presents the results for the static node deployments for the field simulation framework. Note that Figure 4.1a (where $N = 0$) is only used to illustrate the manually deployed static nodes ($SN_{food} \cup SN_{water} \cup SN_{corner}$). The figure is

presented using the surface height data (*meters*) because the algorithm utilised elevation information from the South Esk 'modelled' data set during the optimisation.

In general, based on visual cues, the optimisation performed as expected in that it captured the 'extreme value' of the RoI's topography. Based on Figure 4.1, this phenomenon can be observed from the fact that one hive is located at the middle part (white-brownish) of the RoI that indicates a high elevation data (Figure 4.1b). The algorithm also performed satisfactorily in terms of the sparsity of the entire network (i.e. no extremely-closely located nodes observed).

Evaluating Spatial Interpolation Algorithms

The main focus of this simulation is to compare the performance of three spatial interpolation algorithms (OK, IDW, and SF) based on three factors: computational efficiency, statistical error, and visual aspect. The results in this section are used to justify: based on significant characteristics of the interpolators being compared, which interpolation technique can be seen as a 'balanced' method and is potentially suitable to be considered for the spatio-temporal interpolation algorithm.

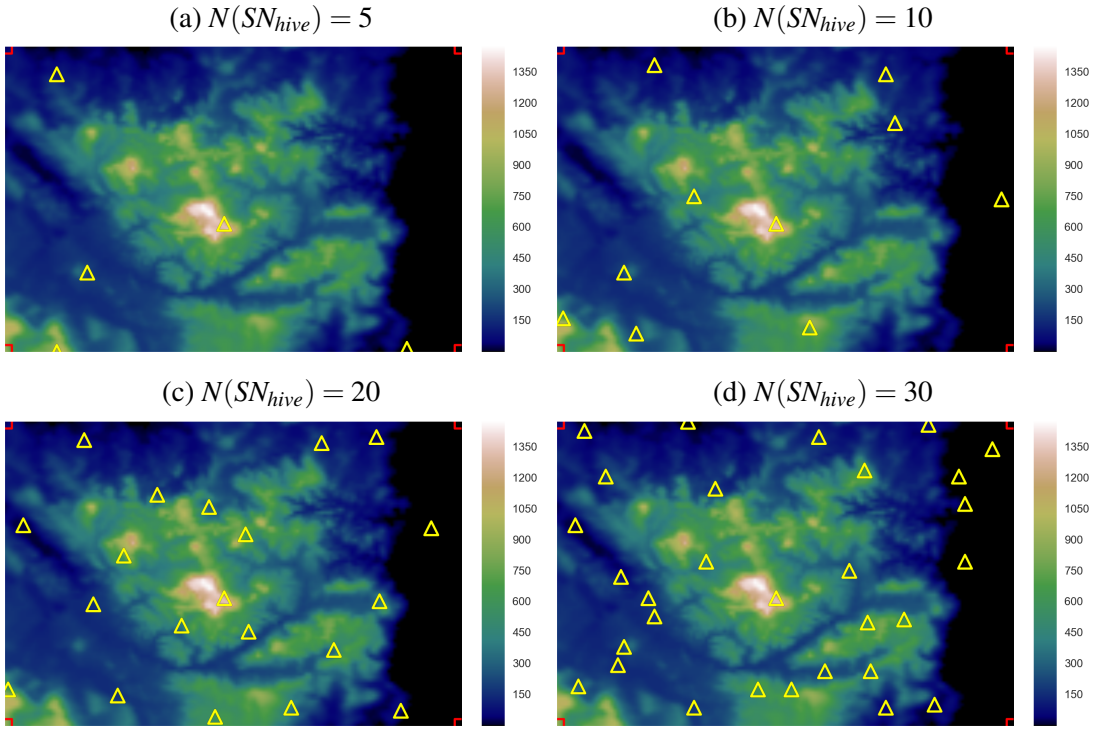


Fig. 4.2 Visual examples for the optimised sensor nodes (yellow triangles) within the RoI bounded by the map's corners (red squares).

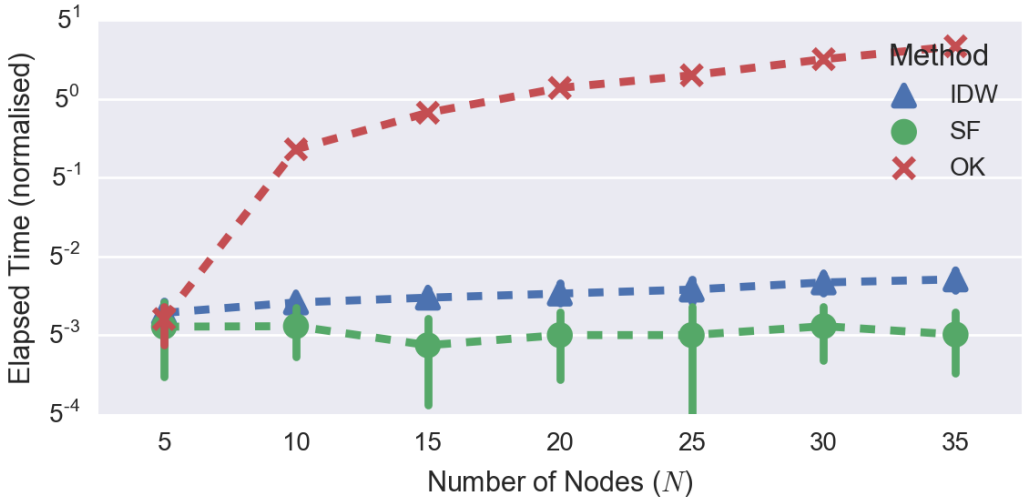


Fig. 4.3 Computational efficiency comparisons between distinct methods. The figure demonstrates the mean and its corresponding 95% confidence interval (vertical bar).

Figure 4.2 demonstrates the weather stations' locations after the EA-assisted ESN optimisation as proposed in Section 3.2. The results of the EA may be different in distinct iterations, therefore it is important to execute the algorithm for a number of replications. In this case, a total of 10 runs were employed in order to have confidence in the results and interpretation [10].

Figure 4.3 illustrates the time elapsed to execute different interpolation methods. The main objective of this computational efficiency test is to observe how the increase in number of observation (sample data) would influence the required time for the interpolator's execution. Therefore, the result is normalised using $x = x - min$, where min is the minimum elapsed time corresponded to different methods. Based on the figure, it is clear that the required computation time of OK grows significantly as the number of nodes increases. This indicates that OK is definitely not preferred if the number of observations is high. On the other hand, SF is the better option for a large-scale problem compared to IDW. Also the elapsed time of IDW increases gradually, while SF possess the advantage of being computationally efficient regardless of the number of nodes (sample data).

In terms of the statistical error comparison, the overall performance of the interpolators yields similar a trend in that the error decreases as the number of nodes increases. Comparing the different methods, OK performed the best (lowest RMSE in Figure 4.4 overall) while SF and IDW have comparable quality. These results confirmed the finding that a geostatistical-based method is more likely to perform better than a deterministic method, such as, IDW and SF in this case [69].

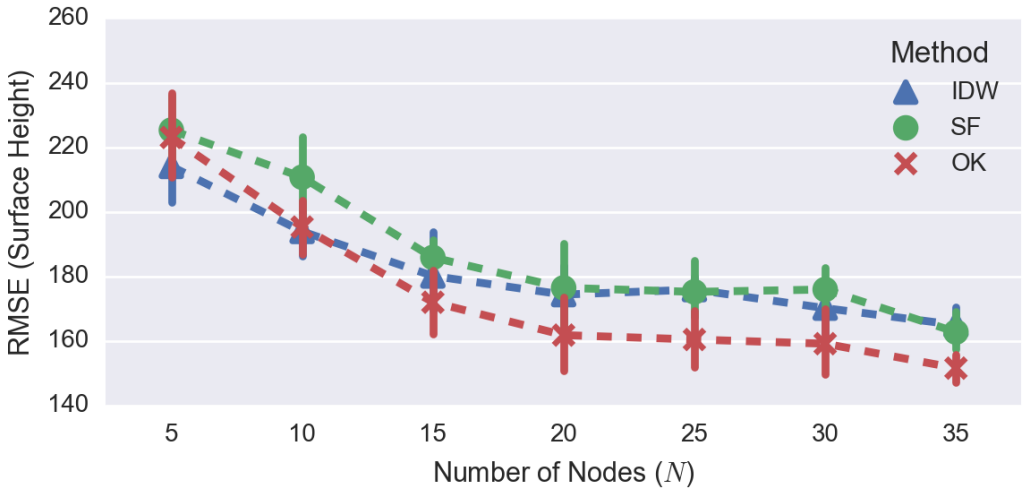


Fig. 4.4 RMSE comparison of elevation data between different interpolators.

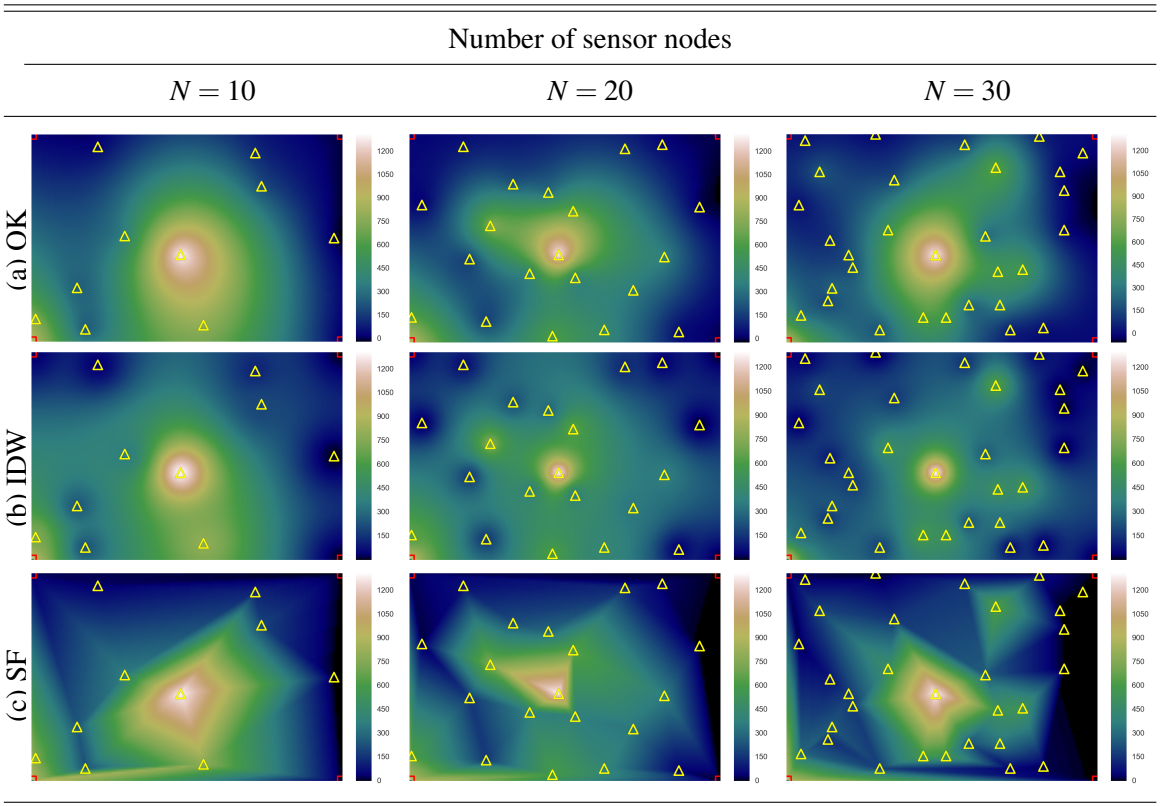


Fig. 4.5 Visual assessment of different interpolators using the optimised sensor nodes and the interpolated/estimated surface height data based on the design shown in Figure 4.2. Each row represents a different method and each column a different number of nodes.

The final assessment is to determine the most realistic interpolation method using visual cues. To do this, the simulation modelled the environmental situation using distinct interpolation techniques within the RoI (Figure 4.5). In general, SF is not considered for environmental interpolation modelling purposes because produces an abrupt result which does not comply with environmental properties in reality (i.e. temperature). Comparing OK and IDW indicates that OK is the preferred method because IDW has a more observable "bull's eye" effect. For instance, based the dark blue area (i.e. sea level) on the right-side of the RoI, OK performed than IDW because IDW produced a noticeable "bull's eye" in that area which is not very realistic.

4.2 Data-driven Bee Behavioural Modelling

This simulation utilises the 'experimental' data set obtained from the bee experiment conducted at Geeveston, Tasmania, which records the bee detections information (Section 3.1.2). The motivation for this bee behavioural statistical modelling is to overcome one of the main disadvantages of RFID systems for insect tracking studies - missed readings. Then, by using pre-defined assumptions and the assistance of computational optimisation, statistical modelling (i.e. inferring) of bee behaviour at time-of-day is possible at the colony-level with a reasonable confidence level.

4.2.1 Experimental Setup

The experimental procedures that will be discussed are based on Section 3.3. Firstly, the 'experimental' data set is grouped on a daily basis for each individual, followed by bee behaviour classification (Section 3.3.1). Then, using the 'processed experimental' data, curve-fitting using an Evolutionary Algorithm is implemented to optimise the Gaussian Probability Distribution Function (PDF) parameters of distinct bee behaviours (Section 3.3.2).

The following sub-section presents the results (i.e. data analytics and discussions) obtained from this simulation.

4.2.2 Results

Bee Detections Data and Bee Activities Distribution

First of all, Figure 4.6 presents the histogram of the unprocessed raw detections data on a daily basis. The figure includes the so-called 'background data' with a *Count* values of approximately 10,000. This results from the fact that the raw bee detections data include readings from dead bees who are within the reader's detection range, thus producing large numbers of very high-frequency detections for each dead bee.

By applying the bee behavioural rules outlined in Section 3.3.1, bee behaviour distribution data can be obtained (Figure 4.7) and will be used for the optimisation process. The figure displays different bee behaviours data (i.e. D_{BTE}, D_{SM}, and D_{FG}) that are subject to the curve fitting process. By comparing Figure 4.6 and Figure 4.7, the 'background data' has been removed. The disappearance of the 'background data' is resulted from the fact that, based on Table 3.5, BTE durations of more than 30*min* are seen as invalid data and will be omitted from the modelling process.

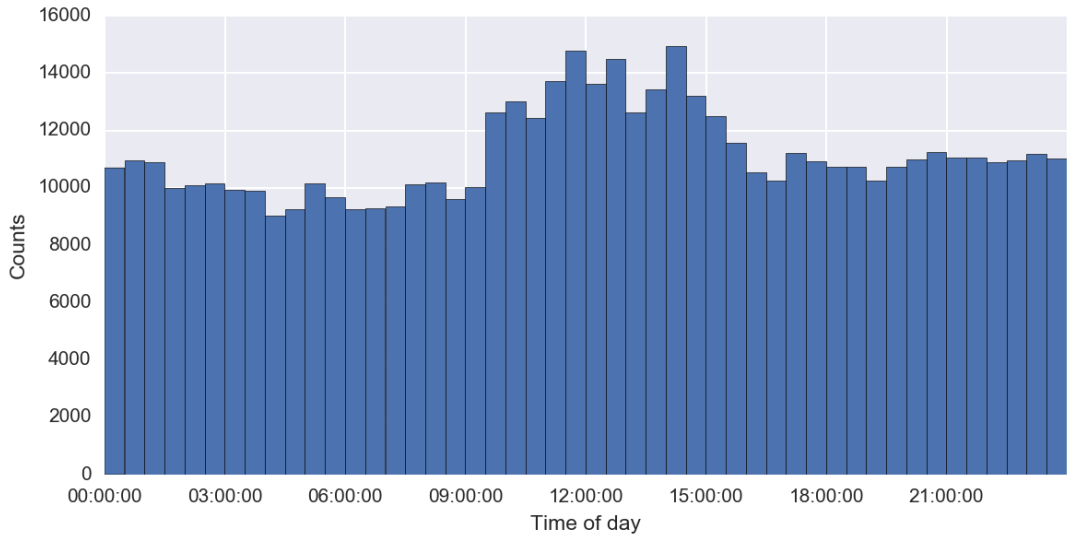


Fig. 4.6 Raw daily detections data from the bee experiment at Geeveston, Australia.

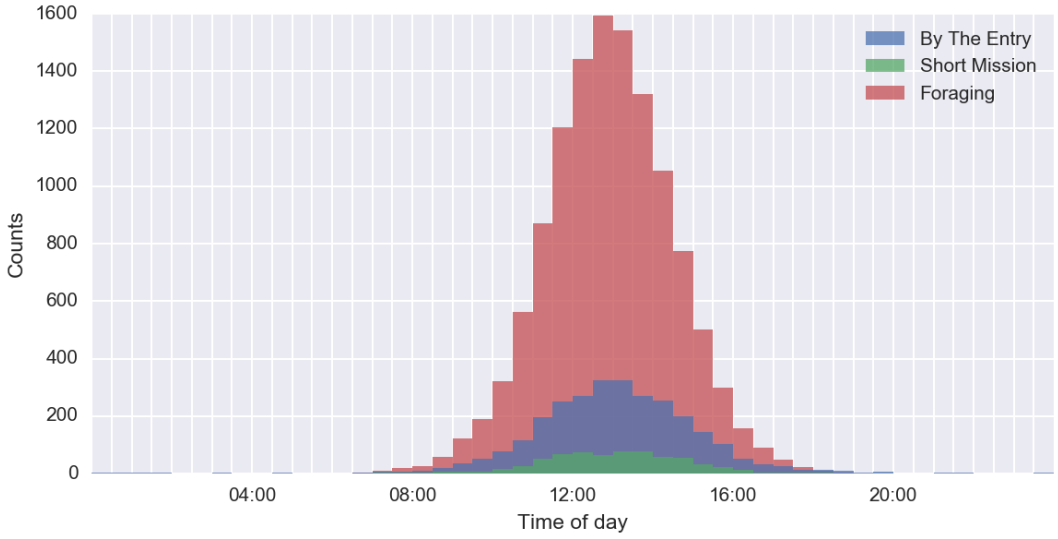
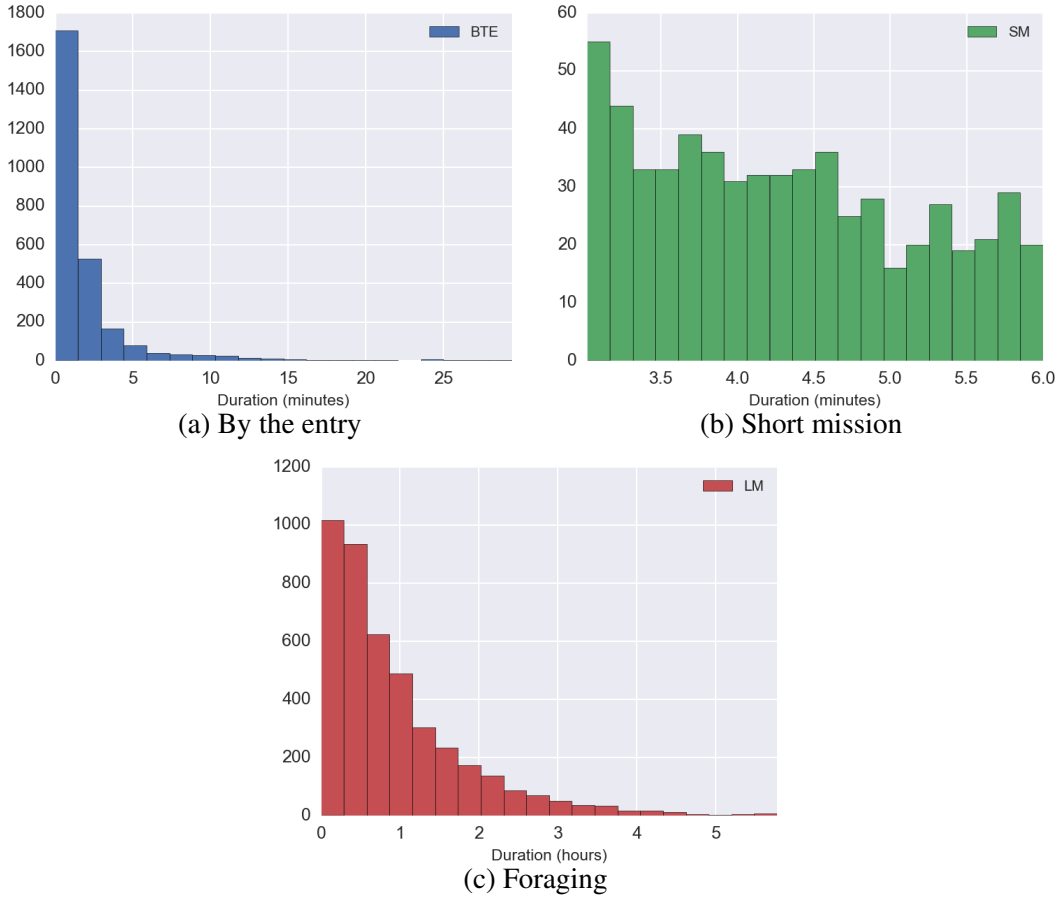


Fig. 4.7 Bee behaviour data after applying the bee classification rules as described in Section 3.3.1. These data will be used for the curve fitting optimisation process to be applied later.

In addition, after the application of the bee behavioural rules to the bee detections data, Figure 4.8 illustrates the activity duration for different activities at the colony-level. Here, we can say that the data set (i.e. for the curve fitting process) is valid because it complies with the pre-defined assumptions listed in Table 3.5. For example, the BTE duration is between 0*min* and 30*min*; while, FG is between 0*hr* and 6*hr*.



(a) By the entry (b) Short mission

(c) Foraging

Fig. 4.8 Bee activities duration from the data in Figure 4.7.

Statistical Inference

At this stage, the curve fitting process (described in Section 3.3.2) is executed on the dataset in Figure 4.7, and the results are shown in Table 4.2 (Gaussian PDF parameters details), Figure 4.9 (data and the curve fitted Gaussian PDF), and Figure 4.10 (percentage of bee behaviour relative to time of day).

Based on Table 4.2, the *BKG* for FG is 0 since it was assumed that an individual bee does not forage earlier/later than sunrise/sunset. Overall, the model suggests that the bees

		Gaussian PDF Parameter				$Area(\%)$
		BKG	I	T_μ	T_σ	
G_{ALL}	G_{BTE}	1.18	322.08	12:53	1h 41m	17.94
	G_{SM}	0.29	77.17	12:57	1h 38m	4.18
	G_{FG}	0.00	1550.91	12:46	1h 33m	77.87

Table 4.2 Parameter details for the curve fitting results shown in Figure 4.9. The last column '$Area(\%)$' indicates the percentage of bees, at a colony-level, involved in a particular behaviour within a day.

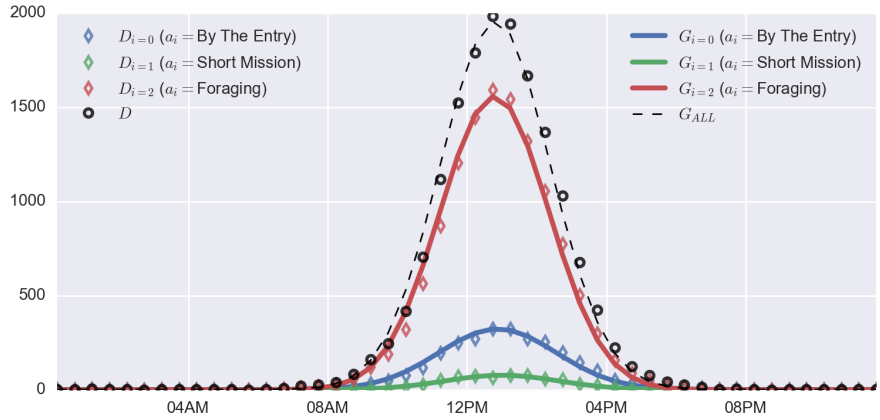


Fig. 4.9 Outcome of the curve fitting process for different bee behaviours. The dots represents data and the solid-line denotes the curve fitted Gaussian PDF. Note that the black dots and black dashed-line denote the summation of data (D) and Gaussian PDF (G_{ALL}), respectively.

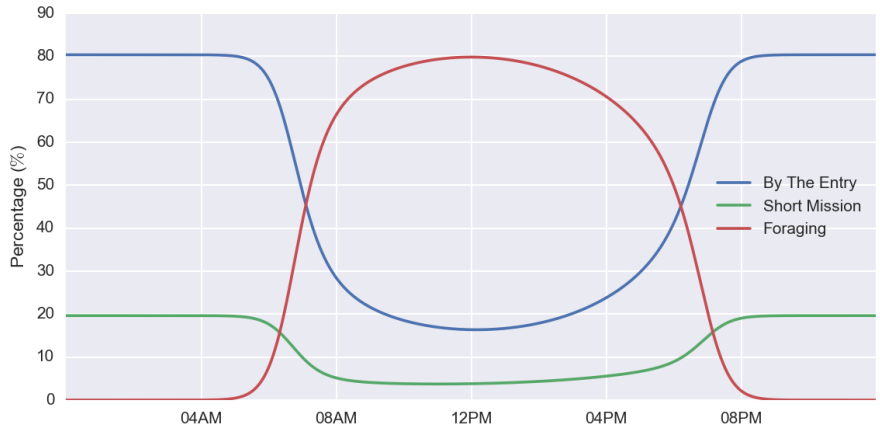


Fig. 4.10 Normalised values (presented as percentages), based on Figure 4.9, of bees involved in different activities relative to the time-of-day.

are mostly active around 12:52pm (average of T_μ) and each of the distributions has a spread of approximately 1h 37m (average of T_σ). In order to quantify the percentage of workers (at colony level) involved in different behaviours within a day, we can estimate them by calculating the 'area under the curve' of each Gaussian PDF. The outcome (last column of the table) suggests that approximately 78% of the workers are engaged in foraging activities, followed by BTE behaviour ($\approx$ 18%), and SM in which only $\approx$ 4% of the colony workers are engaged.

Figure 4.10 presents the normalised proportions based on the results in Figure 4.9. As illustrated, relative to the time of day, the probability of bees (within the colony) involved in different behaviours will vary. The model suggests that bees start forage as early as at 5*am* and ends at 8*pm* (i.e. do not forage before/after sunrise/sunset); during very early morning and late night, approximately 80% of the bees are involved in BTE (e.g. defence) and 20% on short missions near their nest.

Lastly, this simulation provides important information with which to perform a data-driven artificial bee (mobile nodes) simulation described in Section 4.3. The Gaussian PDFs (for different bee behaviours) obtained in this section (Figure 4.10) will be used for the generation of artificial bee data for high-resolution data sampling (Section 4.3.1), so that the simulation is not conducted randomly without any physical basis.

4.3 Swarm Sensing Field Simulation

This experiment involves execution of the 'field simulation' for the South Esk region of Tasmania and the configuration is shown in Table 3.2. In summary, a number of static nodes (acting as weather stations) are deployed to collect hourly environmental data (food/water sources, weather stations at map's corners, and bee hives). Also, for the purpose of this work, I simulated a total of 20 bees each day (mobile nodes) from different hives resulting with a total of 100 bees (5 hives $\times$ 20 bees) in a day, sensing the environmental situation in high frequency (Section 3.4) according to the simulated bee behaviour as described in Section 3.3.

The main objective of this experiment is to mimic the environmental situation in reality and to develop a computational simulation for the CSIRO's Swarm Sensing Project; and eventually, to generate high resolution data sampling within the area under study for environmental modelling purposes (Section 4.4). The following sub-sections will present the field simulation results, including examples of the raw data output and visual demonstrations.

4.3.1 High-Resolution Data Sampling

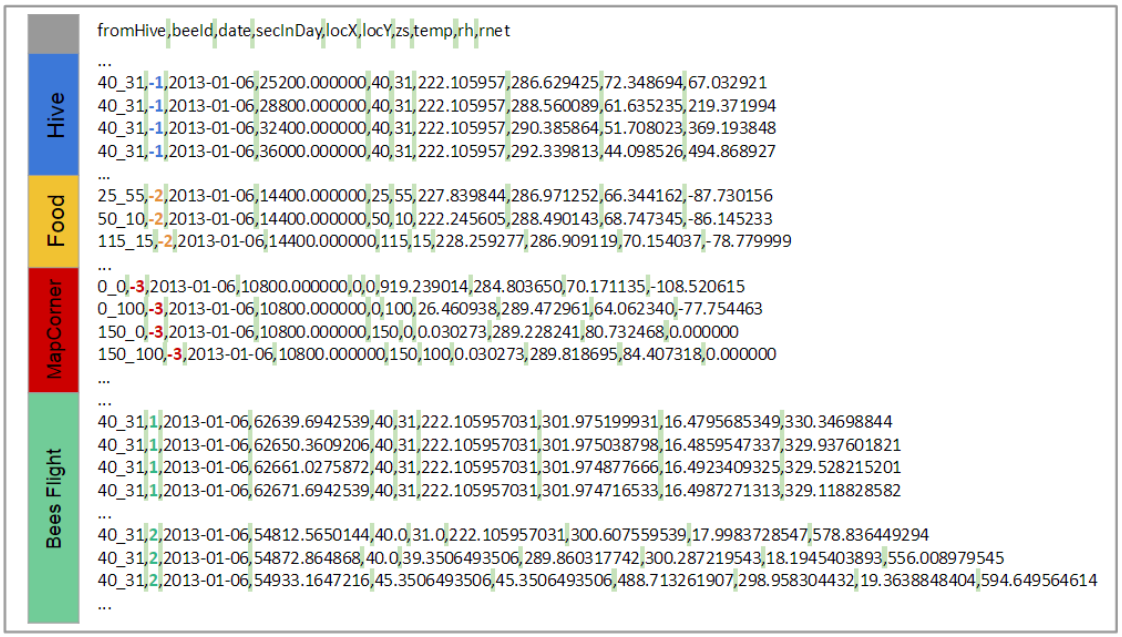


Fig. 4.11 An illustration of the sampled data (CSV output file) from the swarm sensing field simulation. The words with a coloured background on the left are acting as a 'legend' to illustrate the type of sensor node data in 'beeId' column. Also, the *commas* are highlighted in light-green for ease of visualisation.

Figure 4.11 presents the high resolution generated comma-separated-values (CSV) data produced by the field simulation. To illustrate, a 'legend' on the left is used with the colour corresponding to the sensor node type in the *beeId* column: hive (blue, $beeId = -1$), food source (yellow, $beeId = -2$), weather stations at map's corners (red, $beeId = -3$), and the actual bee data (green, $beeId \geq 0$) that 'sense' the environment while they fly.

4.3.2 Data Visualisation

This sub-section will present the high-resolution data generated. In this part, results will be presented for a particular simulated bee over one day from each hive. The hourly data from the static nodes will also be presented throughout the day. Results are displayed in Figure 4.12 and Figure 4.13.

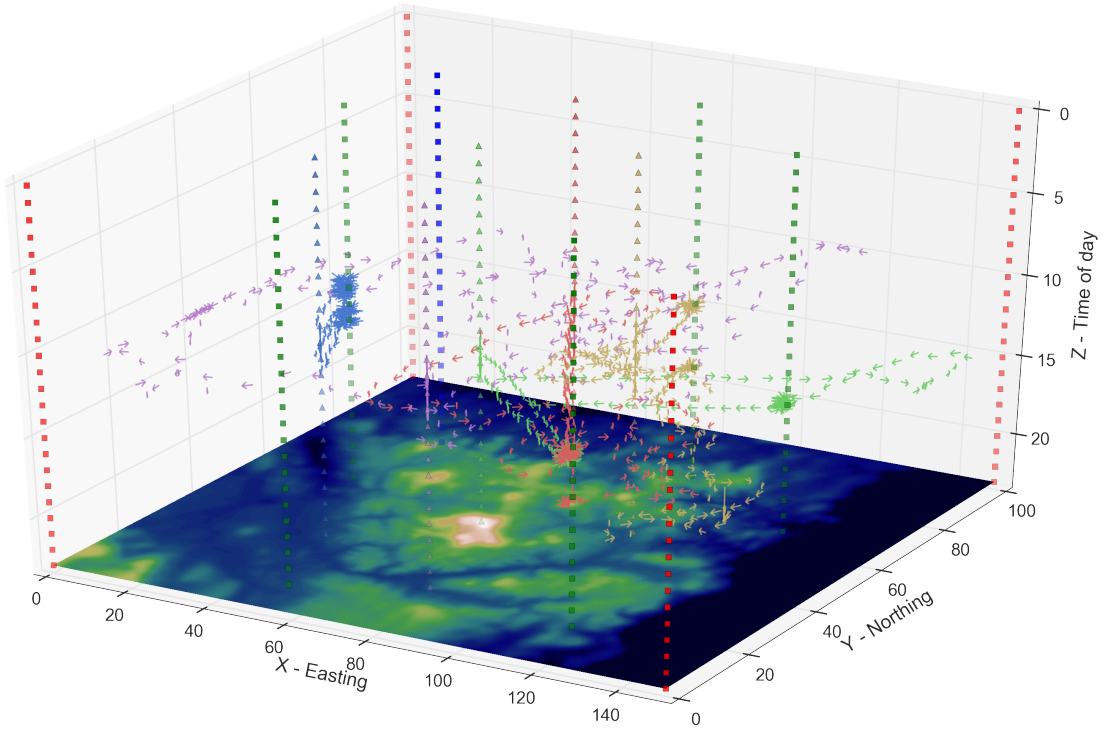


Fig. 4.12 A demonstration of the field simulation, showing the data collected within the RoI (X and Y spatial dimension) throughout the day (Z - time of day). Squares denote hourly data detections from the static sensor nodes: food (green) and/or water (blue) sources, the weather stations at the map's corners (red). Triangles are the hives that are represented using different colours. Finally, the arrows represent bees' flight paths and each of them is a particular datum obtained by 'sensing' the environment. Flight paths of a particular colour match the colour of the hive from which that bee originated.

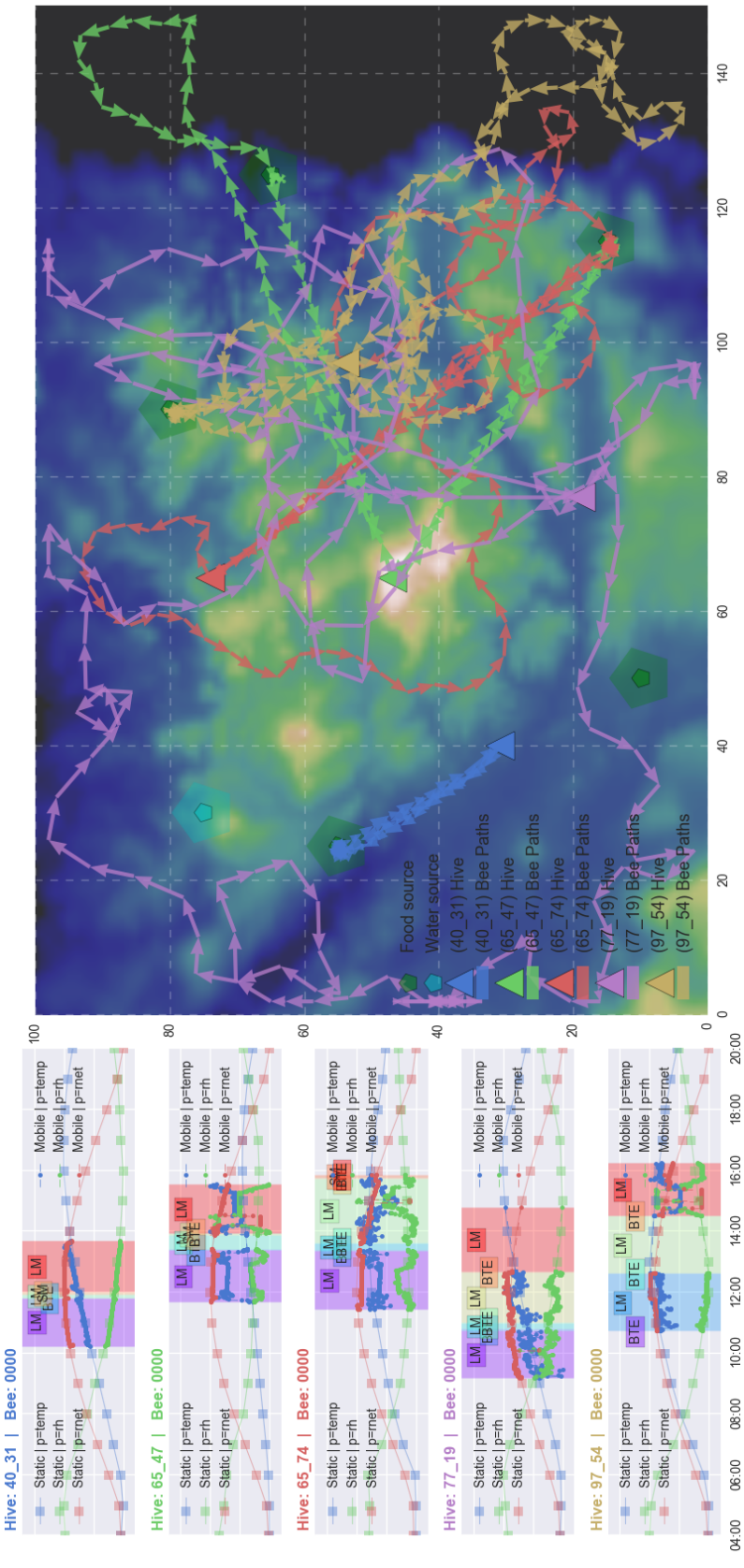


Fig. 4.13 A dashboard illustrating the data obtained from the proposed field simulation framework. The left pane represents the time-line throughout the day for a single bee simulated from distinct hives, and includes the following components: (a) hourly data collected from each hive. (b) simulated bee activity in that day and its corresponding duration; (c) high-resolution data generated by 'sensing' the environment from mobile sensor nodes. The right pane is a top-down view based on Figure 4.12, that disregards the temporal dimension (time-of-day).

4.4 Environmental Modelling

The final simulation is to utilise the simulated data/observations (Section 4.3) within the RoI and to generate a near-to-reality environmental model for the Swarm Sensing project conducted by CSIRO. To achieve this, the Spatio-temporal Interpolation (STI) procedure described in Section 3.5 will be employed. This experiment will execute the framework that has been developed and demonstrates the results obtained from the implementation. Discussions of the outcome will also be provided.

4.4.1 Spatio-temporal Variogram Model

The initial effort is to generate the spatio-temporal variogram to model the space-time interaction of the residual component within the observations as discussed in Section 3.5.1. Figure 4.14 illustrates a particular daily spatio-temporal variogram which is used for the weighting mechanism in the STI algorithm. However, since the model is generated on a daily basis, the simulation will produce slightly different variogram parameters (i.e. range and sill) for different days. Such a phenomenon is acceptable in a way that environmental situation in reality might possess a non-identical variability at distinct days.

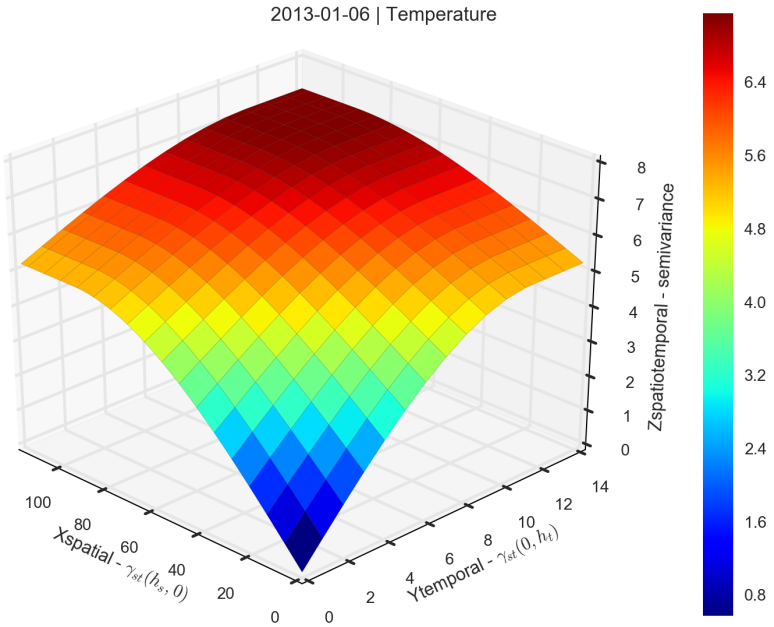


Fig. 4.14 Spatio-temporal empirical variogram model generated using the static-only nodes for the temperature data on 06 January 2013.

4.4.2 Data Pre-processing

Prior to the STI algorithm execution, pre-processing of the high-resolution observations is needed (Section 3.5.2). At this stage, the high-resolution spatio-temporal simulated data (Figure 4.15) will used in a way to generate a 3D 'cube-like' data set that holds the mean (μ) and standard deviation (σ) values resulting from observations within the individual partitions. The output is illustrated in Figure 4.16.

Based on Figure 4.16, the x and y axes are the spatial dimensions and are sub-partitioned into 11 and 12 partitions respectively; while the z-axis represents the temporal unit (time-of-day) that is divided on an hourly basis leading to a total of 23 sub-partitions within a day. For visualisation purposes, each temporal index can be distinguished using distinct colour. For instance, the 2D-spatial 'processed' data (dots) are in purple at 1am (top) and gradually change colour to red as they progress through the day to 11pm (bottom). As previously mentioned, each dot represents a datum that includes a μ value (for the actual estimation) and a σ value (for the corresponding error measurement), to be used within the STI algorithm at a later stage.

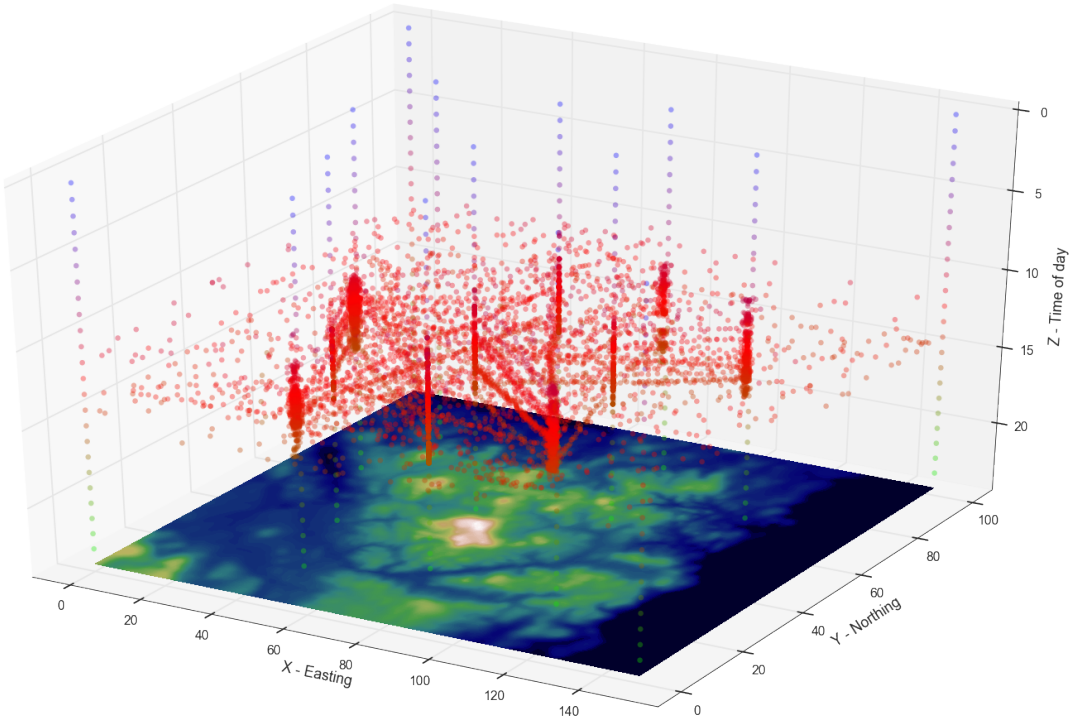


Fig. 4.15 High resolution data obtained from the Swarm Sensing hybrid sensor network (Section 4.3). Each dot is a data point 'sensed' by either a fixed or a mobile node. The data are denoted using different colours based on time-of-day (z-axis): early morning (00:00, blue); noon (12:00, red); and late night (24:00, green).

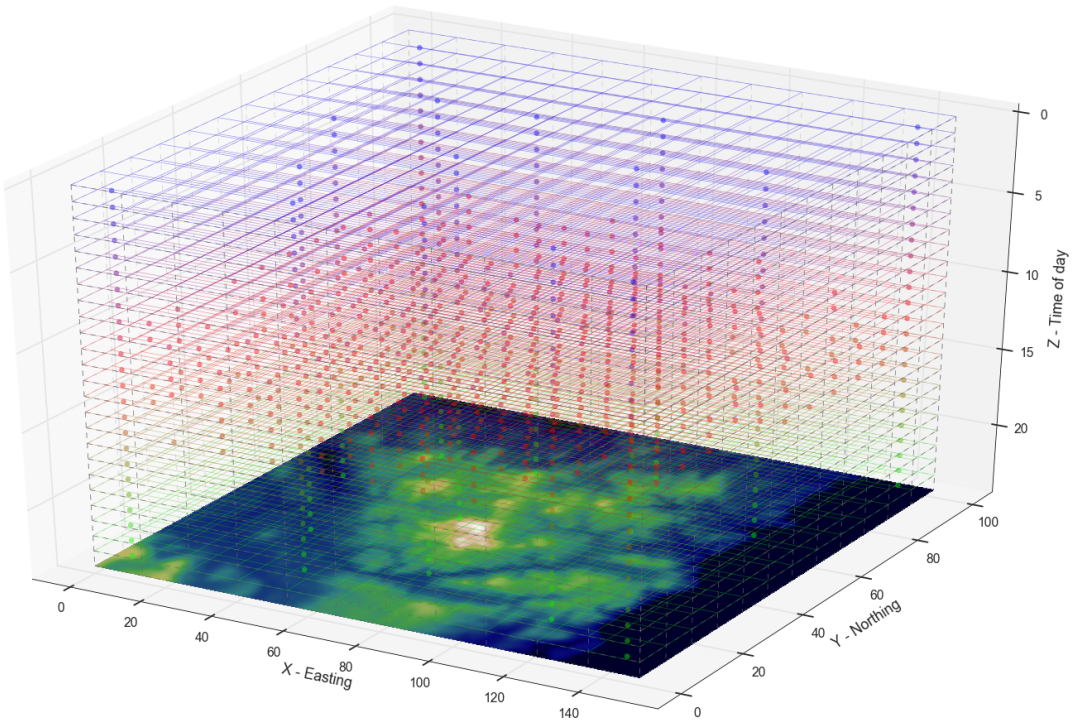


Fig. 4.16 Visualisation of the data set after the 'pre-processing' procedure. Each dot representing data holds the mean (μ) and the standard deviation (σ) value that will be used for the interpolator's estimation and its corresponding error, respectively.

4.4.3 STI Assessment

This experiment validates the performance of the proposed STI-algorithm based on a 'benchmark' data set (Section 3.3) that consists of six sites as described in Table 3.4. In addition, the 'modelled' data set will also be used for validation because the data was originally used for the field simulation's data sampling process. Furthermore, please note that since the site locations are not regularly gridded as shown in the 'coordinates' column in Table 3.4, Equation 3.16 will be employed for the value estimation at different sites over a specified time-frame. Finally, after the STI-algorithm execution, an 'estimated' data set will also be produced for final validation purposes.

From the experimental configuration discussed previously, three data sets will be used for comparison and validation purposes. As such, the following *Schemes* will be used to identify comparisons between paired data sets: (a) *Scheme 1*: 'modelled' and 'estimated'; (b) *Scheme 2*: 'modelled' and 'benchmark'; (c) *Scheme 3*: 'estimated' and 'benchmark'. Lastly, the validation will be carried out using three different error measurements, they are: the Pearson's *r* correlation of coefficient (Equation 3.27), the mean absolute error (Equation

	Site #	Pearson *r*			MAE			RMSE		
		temp	rh	rnet	temp	rh	rnet	temp	rh	rnet
(a) Scheme 1	1	0.97	-	-	0.88	-	-	1.37	-	-
	2	0.98	0.97	-	1.76	4.81	-	1.99	6.92	-
	3	0.98	0.99	-	0.69	2.65	-	0.95	3.49	-
	4	0.97	-	-	1.21	-	-	1.43	-	-
	5	0.96	0.93	-	1.23	6.80	-	1.63	8.75	-
	6	0.99	0.97	1.00	0.46	2.96	16.30	0.64	4.56	25.67
(b) Scheme 2	1	0.94	-	-	1.43	-	-	1.88	-	-
	2	0.92	0.78	-	1.90	16.06	-	2.41	20.43	-
	3	0.94	0.82	-	1.78	9.29	-	2.29	12.55	-
	4	0.87	-	-	1.78	-	-	2.41	-	-
	5	0.95	0.79	-	1.62	11.83	-	2.08	15.55	-
	6	0.93	0.85	0.91	1.61	9.02	97.93	2.04	12.57	123.63
(c) Scheme 3	1	0.91	-	-	1.95	-	-	2.56	-	-
	2	0.93	0.81	-	3.40	18.08	-	3.75	21.65	-
	3	0.93	0.83	-	1.69	8.90	-	2.17	11.94	-
	4	0.89	-	-	2.27	-	-	2.80	-	-
	5	0.92	0.78	-	1.59	10.68	-	2.04	13.94	-
	6	0.94	0.84	0.91	1.49	8.96	95.43	1.91	12.41	122.63

Table 4.3 Summary error statistics of the STI-algorithm.

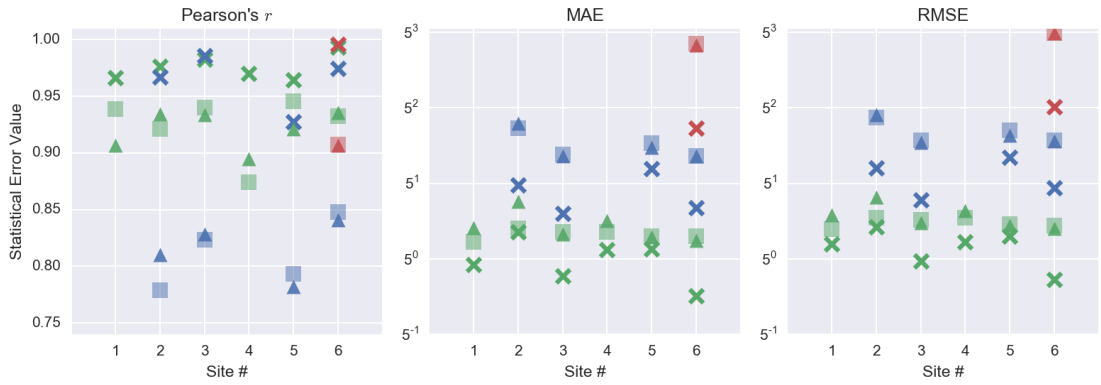


Fig. 4.17 Scatter plot based on Table 4.3: x-axis denotes the sites, and y-axis is the error values for corresponding error measurements (i.e. Pearson's *r*, MAE, and RMSE). The shapes are used to distinguish the *Scheme*: square (*Scheme 1*), triangle (*Scheme 2*) and cross (*Scheme 3*); The colours represent distinct parameters: temperature (green), relative humidity (blue), and solar radiation (red).

3.26), and the root mean squared error (Equation 3.25). The results are presented in Table 4.3 and Figure 4.17.

Scheme 1: 'modelled' and 'estimated' data sets. First, the performance of the proposed STI algorithm output ('estimated' data) against the South Esk 'modelled' data set is assessed. The results show that the proposed STI-algorithm provides $r \geq 0.94$ and, more interestingly, $r \approx 1.0$ at Site 6 (Table 4.3 and Figure 4.18). This is because, based on Figure 4.19, we observe that the surroundings around Site 6 possess low variance. Also, the figure shows that the proposed STI algorithm is able to capture the overall surface structure of the area under study. Therefore, the proposed technique is able to provide a reasonable level of accuracy (i.e. both statistically and visually) and is suitable for environmental modelling applications.

Scheme 2: 'benchmark' and 'modelled' data sets. Then, we introduce the 'benchmark' data set (Section 3.1.4) for validation purposes to assess the quality of the South Esk 'modelled' data set, which was used for the entire Swarm Sensing field simulation. The results in Table 4.3 show that the 'modelled' data set is acceptable, with $r \geq 0.88$ on average.

Scheme 3: 'benchmark' and 'estimated' data sets. Finally, we conduct a comparison between the 'benchmark' data set and 'estimated' data set. Such a process can be seen as a performance evaluation of the entire Swarm Sensing field simulation, because the 'benchmark' data set and the 'estimated' data set were obtained/processed using entirely different approaches (e.g. data sampling and application of STI algorithm was involved with the 'estimated' data set). Based on such assumptions, it is expected that the errors in Scheme 3 would approximate the combined error of Scheme 1 and Scheme 2 (i.e. r in Scheme 3 is less than either Scheme 1 or 2); Yet, interestingly, there is an anomaly within temperature where r in Scheme 3 is higher than that of Scheme 2 for Sites 4 and 6.

Finally, the MAE and RMSE error measurements within Table 4.3 and Figure 4.17 is investigated. Similar to the previous discussion, it is shown that Scheme 1 provides better results (i.e. lower statistical error) than those of Schemes 2 and 3. Nevertheless, an anomaly occurred at Site 3 where the temperature RMSE of Scheme 1 is higher than that of Scheme 2. This phenomenon can be explained by the significant change of topography (and temperature data, which is highly correlated with elevation) around Site 1 resulting from the post-processing of the spatiotemporal data before the STI algorithm execution. In this case, based on the right of Figure 4.19, the location of Site 1 is just outside of the 'partitioned' area around it, leading to a higher discrepancy between the 'modelled' and the 'estimated' values.

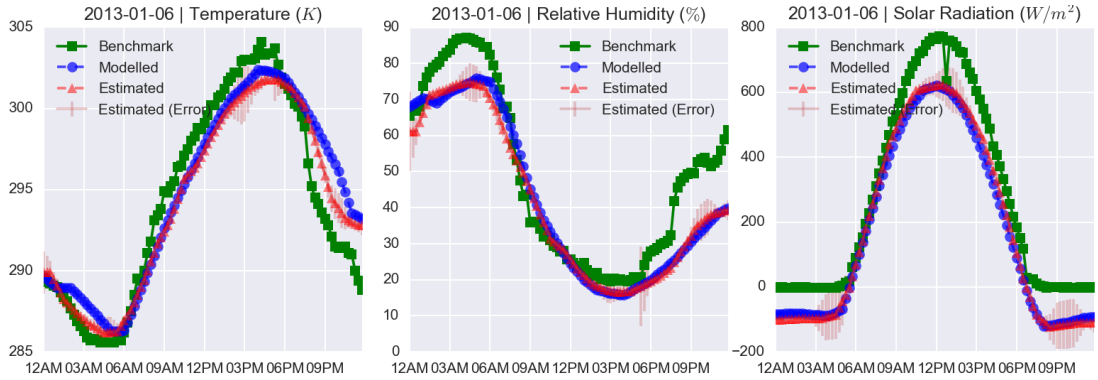


Fig. 4.18 Timeline plot showing the values from three different data sets: 'benchmark' (green), 'modelled' (blue), and 'estimated' (red). This example is based on Site 1 (Ben Lomond) on 06 January 2013.

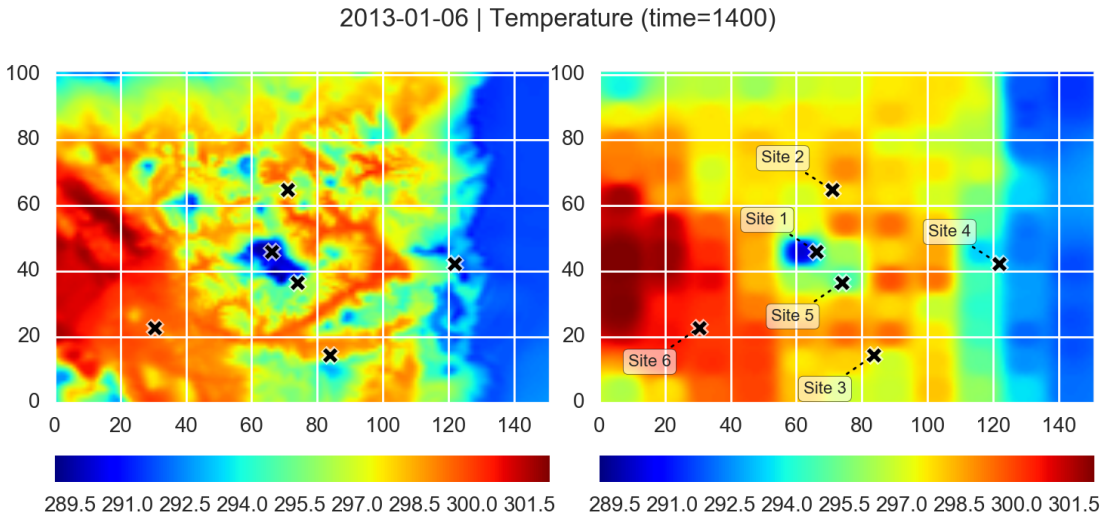


Fig. 4.19 A visualisation demonstrating the 'modelled' (left) and 'estimated' (right) temperature data on 06 January 2013. It also presents the spatial locations of six weather stations (denoted using '×') from the 'benchmark' data set.

4.4.4 High Resolution Environmental Modelling

Lastly, this simulation is executed to demonstrate that the proposed STI-algorithm is capable of generating a spatio-temporal environmental model that is realistic, in that it produces smooth transitions within the spatial surface and along the temporal dimension. In this particular example, temperature data are employed for the high-resolution environmental for the day on 06 January 2013.

The result is shown in Figure 4.20. Each row in the figure illustrates the outputs generated from different time frames and intervals on that day. It is shown that the algorithm can produce smooth transitions on an hourly basis (Figure 4.20a and Figure 4.20b) and also in 10 minute intervals (Figure 4.20c). The technique is also capable of estimating the values at even higher resolutions, such as the irregularly-spaced spatial locations and time points.

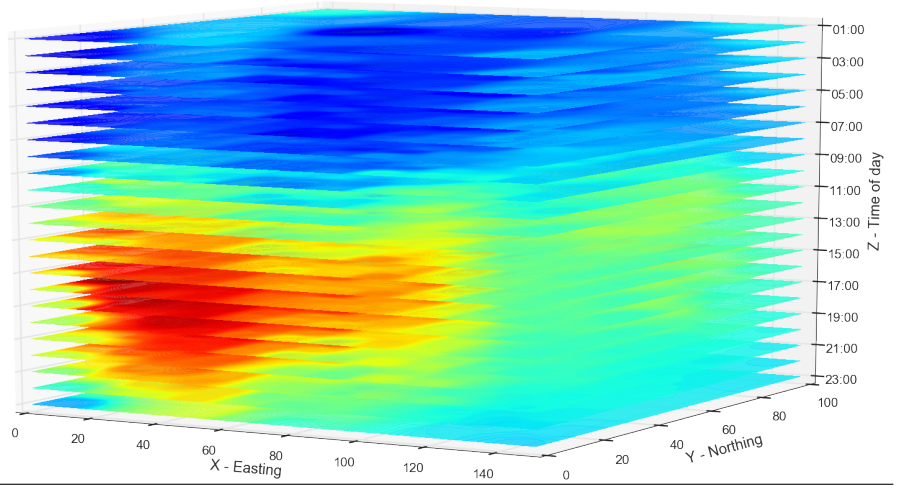
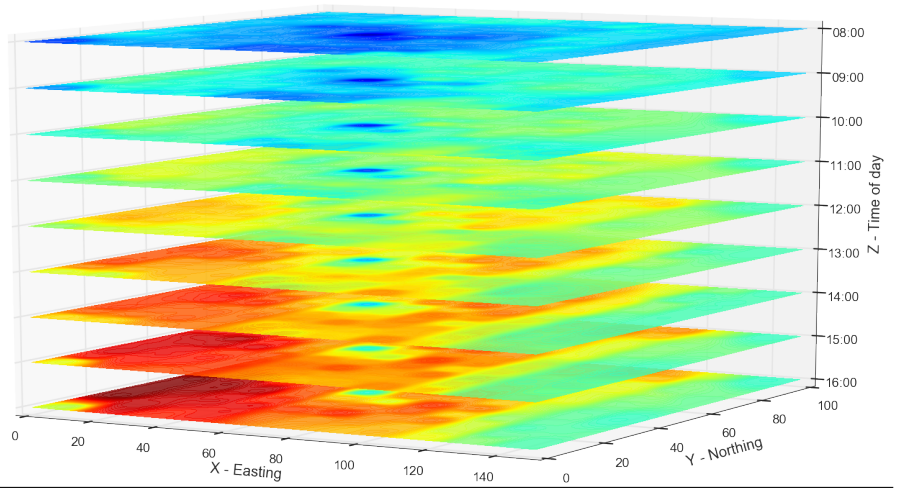
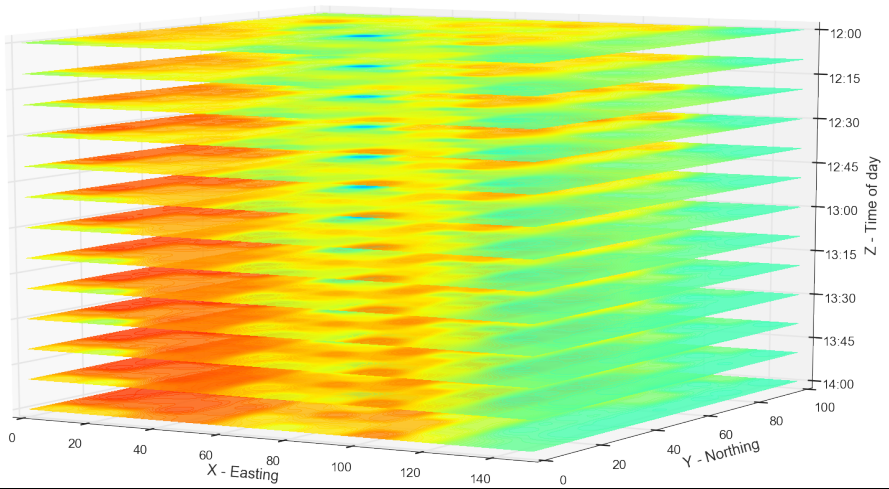
Time-of-day and Interval	Visualisation
(a) 1am $\cdots$ 11pm $interval = 1h$	
(b) 8am $\cdots$ 4pm $interval = 1h$	
(c) 12pm $\cdots$ 2pm $interval = 10m$	

Fig. 4.20 Demonstration of high-resolution spatio-temporal environmental modelling. The x and y are the spatial dimensions, namely, easting and northing respectively; whilst, the z is the time-of-day with different intervals. For instance, (a) and (b) are generated in hourly basis, and (c) is produced in 10 minutes interval.

Chapter 5

Discussion

This chapter provides a comprehensive discussion of the implementation results presented in Chapter 4. The following sections discuss different components of the framework and mainly focusing on the folowing aspects where appropriate: (i) analysis of the results; (ii) fundamental assumptions of the simulation; and (iii) the limitation of the method or simulation.

5.1 Spatial Sampling of Static Node

To begin with, after the manual deployments of fixed sensor nodes (weather stations, food and water sources), this work utilised an evolutionary algorithm to optimise the hive locations so that the network can to capture the *representativeness* of the entire region using a pre-defined number of nodes (Section 4.1). In this work, the *representativeness* of a network is calculated based on the error between the original data set and the interpolated estimation using the optimised hive locations.

There is no doubt that an increase in the number of sample data will result in better understanding of the environment (surface height of the RoI in this experiment). Nevertheless, it is also worthwhile to consider the cost to set up and maintain the sensor networks, and eventually, to determine whether a high number of nodes is necessary to obtain a certain level of understanding of the RoI for a particular purpose. However, these issues are outside of the scope of this work.

An experimental simulation, using the proposed spatial sampling algorithm, was conducted to assess different spatial interpolations (Section 4.1.2). Table 5.1 summarises the comparison of the three different interpolators; OK, IDW and SF. The table is generated based on the requirements of the Swarm Sensing application where a large number of observations will be incorporated into the spatio-temporal interpolation process. Firstly, since

Item	Method		
	OK	IDW	SF
Computational Efficiency	-1	0	+1
Statistical Error	+1	0	0
Visual Assessment	+1	0	-1

Table 5.1 Summary comparison of interpolators. The values are to be interpreted as: +1 (most preferred), 0 (neutral), and -1 (least preferred).

computational efficiency is the most critical factor for a swarm sensing application (e.g., processing highly-dense observations), a Kriging-based technique is not suitable because it very computationally demanding. IDW and SF exhibit similar performance as far as statistical error is concerned. It is suggested that an *IDW-based* method is an ideal option for the spatio-temporal interpolation algorithm because SF produces abrupt changes in the surface.

5.2 Data-driven Bee Behavioural Modelling

The bee behaviour classification in this work (Chapter 3.3.1) was interpreted based on the readings frequency (i.e. using threshold values as illustrated in Table 3.5). Based on this and with the additional consideration that missed readings are a major issue in insect tracking using RFID systems, this indicates that higher thresholds have lower confidence levels in the analysis. For example, there is a high certainty level for bees involved in BTE-related activities (i.e. hive defence, micro-climate of the hive, etc) and SM activities (i.e. orientation flights, short flights, or wondering around the hive); FG activities, on the other hand, can only be inferred with a very high uncertainty resulting from the missed readings issue. It is because, if the system misreads one of the readings, it is almost impossible for a biologist to accurately interpret what the individual bee was actually doing at that particular time.

To conclude, the main contribution of this model is to infer bee behaviour at a colony level using incomplete RFID detection data and to use that information to execute an artificial bee simulation utilising the generated Gaussian probability distribution functions. The model estimates the proportion of bees involved in different behaviours (i.e. by the entry, short mission, and foraging) at different times of day (e.g. 10am, 2pm, 10pm). Such an effort is crucial to simulate artificial bee activities over a day using bee activity data obtained from

Abbreviation	Behaviour	Bee activities	Certainty
BTE	By the entry	Hive defence, maintaining the hive micro-climate	High
SM	Short mission	Orientation flight, walking around the hive	Medium
FG	Foraging (in-hive)	Searching or exploiting food/water sources	Low
FG	Foraging (out-of-hive)	Depositing food/water into the hive	Low

Table 5.2 Summary of bee behaviour with the 'possible' activities for each classification. The level of certainty for distinct behaviours is given in the last column (Certainty).

experimental work conducted at Geeveston, Tasmania, thus ensuring that there is a physical basis for the simulation. In swarm sensing field simulation, it is assumed that bees do not forage before sunrise or after sunset.

One of the major limitations of this model is the data quality issue. This was mainly caused by hardware limitations such as misreadings of bee detections and computer failure. In such cases, our data analysis would be erroneous because the estimated duration of a particular bee's activity would be greater than the actual duration.

Domain knowledge is critical to define the 'threshold' configuration (Table 3.5) in addition to considering the bee species, experiment's location (weather condition in that area), etc. In this work, the thresholds are defined based on field observation and a little knowledge on bee behaviour.

Another limitation is that the inability to replicate the bee experimental data obtained from the field. This is because of several reasons: the geographical location of the experiment, the frequency of field visit to tag the bees, and the seasonality effect that has a great influence to bee activities.

5.3 Swarm Sensing Field Simulation

The actual Swarm Sensing field simulation (hybrid ESN) is then used to generate high-resolution observations – data sampling (Section 4.3). The fixed sensor nodes within the

networks collect data on an hourly basis. Then, the framework simulates a number of bee activities throughout the day and, based on the activity that a particular individual bee is involved in, data will be sampled at different frequencies. Also, flight paths will be assigned for bees that are involved in out-of-hive foraging (e.g. scout, recruit, exploit) so that they sense the environment as they fly within the area under study. The sampled spatio-temporal data points are stored in a Comma Separated Values (CSV) data format.

The bee foraging data set used in this work (Section 3.1.3) that was done by one of the team members in Swarm Sensing team at CSIRO. This model was developed based on the information about bee foraging behaviour obtained from the current literature and is subject to further improvement over time. Since this work employed the bee flight paths from this simulator within the field simulation for high-resolution data sampling (Section 3.4), it is also worthwhile to consider the modelled data's credibility to ensure the entire field simulation's validity. However, this area is outside of the scope of this dissertation.

5.4 Environmental Modelling

Lastly, by using the high-resolution sampled data from the field simulation, a number of experimental simulations are executed to assess the proposed Spatio-temporal Interpolation (STI) algorithms (Section 4.4). The method involves a combination of extension and reduction based approaches for a smooth transition between space and time. The results revealed a high correlation between the data (originally used for the simulation) and the estimation (interpolated values), suggesting that the algorithm is statistically acceptable. Moreover, the interpolation technique also demonstrated the ability to estimate values at irregularly-gridded spatial locations and time points. These results indicate that the proposed hybrid approach (extension and reduction) STI algorithm is suitable for high-resolution environmental modelling applications.

This dissertation focuses on the environmental study, thus, such an assumption requires an interpolation method that can produce a smooth continuous surface of the study region to comply with the environmental situation in reality. Based on the result presented in Figure 4.3, it shows that the Shape Function (SF) interpolator is the most efficient algorithm (i.e., elapsed time does not increase significantly as the number of nodes increases). This has justified that, in the case where surface smoothness is not of an issue, shape function could become the most suitable method for high dense data observations. For instance, Li uses a SF-based spatio-temporal interpolation algorithm to estimate the house price [45].

Chapter 6

Conclusion and Future Work

This work mainly focussed on proposing a near to reality field simulation for swarm sensing application to obtain highly-dense data so that an initial data analysis could be undertaken while the micro-sensor development is still ongoing. Data collected from the swarm of mobile nodes (insect in this case) will be analysed to better understand bee behaviour, and eventually, to help discover the environmental impacts that are contributing to the detriment in bee populations worldwide.

6.1 Research Contribution

The main contribution of this thesis is the proposal of a Swarm Sensing field simulation framework and a new variation of the Spatio-temporal Interpolation (STI) technique for high-resolution environmental modelling purposes. The key research component addressed in order to achieve this objective are:

1. *Environmental sensor network deployment – spatial sampling.* An optimisation method is proposed to provide a near-optimal design static sensor networks with minimal historical knowledge within the region of interest. For instance, this work only employed the surface height (i.e. elevation) data to serve the purpose. This method has been used for two purposes in this work: (i) For the Swarm Sensing field simulation to optimise the locations of bee hives; and (ii) To deploy a sensor network, with a pre-defined number of nodes, for an empirical study to determine a 'balanced' spatial interpolation technique that considers computationally efficiency, acceptable statistical error, and the near-to-reality environmental situation.

2. *Data-driven statistical inference of bee behaviour.* This part utilises the real experimental data obtained from the bee experiments being conducted by CSIRO, to model

the bee behaviour within a day. A curve fitting optimisation process is executed to obtain the Probability Density Function (PDF) of the Gaussian distribution for the bee detection data. The main focus of this step is to address following question: what is the probability of a specific bee behaviour at a particular time-of-day? This component is crucial for the artificial bee simulation process within the Swarm Sensing field simulation.

3. *The Swarm Sensing simulation for high-resolution spatio-temporal data sampling.* This is the actual Swarm Sensing field simulation to 'sense' the environment from static (i.e. weather stations, food and water sources, bee hives, etc) and mobile (insects) sensor nodes based on different reading frequencies. For instance, the static weather stations record data on an hourly basis, and the mobile nodes collect environmental data according to the insect's behaviour at high frequency (i.e. seconds).

4. *A Spatio-temporal Interpolation (STI) algorithm for environmental modelling.* This work proposes a hybrid (i.e. extension and reduction) approach to create a high-resolution environmental model from a huge amount of data obtained from the field simulation (Item 3). Firstly, a geo-statistical method is utilised to model the space-time interactions by using only the static nodes for computational efficiency considerations. Then, the entire input data are pre-processed to address the *sampling density* issue. Finally, the hybrid STI algorithm is executed to create the environmental model. This framework can estimate the parameter value at any irregular spatial and temporal location with reasonable accuracy (i.e. low statistical error), providing that the input data is valid.

The author has implemented the framework/software using the Python programming language and a series of experimental simulations have been conducted to demonstrate and validate the results.

6.2 Future Work

One of the useful future research goals would be to develop a novel algorithm to curate the experimental bee data obtained by CSIRO (Section 3.1.2). As with any other statistical model, data quality is a crucial factor to have a certain level of confidence that the developed model is accurate. In this case, the Swarm Sensing simulation model is dependent on the bee experimental dataset, which suffers from the limitation of high a misreading frequency.

Another future research direction is to utilise other data types (e.g., sound recording within the hive, scale, and camera) to infer better the bee behaviour at both individual-bee

and colony level. Such an effort could beneficial to the artificial bee simulation framework proposed in this dissertation (Section 3.3.3).

Scientists could analyse the data obtained from the swarm sensing field simulation generated from this work (Section 4.3.1) for a wide range of applications. For example, utilising analytical method to discover pattern within the sampled data and correlate the 'sensed' environmental data to make inferences of insect behaviour, in which might contribute to the understanding of detreimental impacts of bee population worldwide.

The spatio-temporal interpolation (STI) algorithm proposed in this dissertation is subject to improvement particularly within the space-time interaction model (e.g. variogram). Also, since the performance of any interpolator is highly dependent on the type of data set, the proposed STI algorithm is expected to perform differently in other cases. Thus, it is also worthwhile to apply and compare the proposed STI method in this dissertation with other algorithms in order to verify the conclusion drawn in this work.

References

[1] B. Warneke, M. Last, B. Liebowitz, and K. S. J. Pister. Smart dust: communicating with a cubic-millimeter computer. *Computer*, 34(1):44–51, 2001. doi:10.1109/2.895117.

[2] K. Martinez, J. K. Hart, and R. Ong. Environmental sensor networks. *Computer*, 37(8):50–56, 2004. doi:10.1109/MC.2004.91.

[3] J. K. Hart and K. Martinez. Environmental sensor networks: A revolution in the earth system science? *Earth-Science Reviews*, 78(3–4):177–191, 2006. doi:10.1016/j.earscirev.2006.05.001.

[4] J. H. Porter, E. Nagy, T. K. Kratz, P. Hanson, S. L. Collins, and P. Arzberger. New eyes on the world: Advanced sensors for ecology. *BioScience*, 59(5):385–397, 2009. doi:10.1525/bio.2009.59.5.6.

[5] Chong C. Y. and S. P. Kumar. Sensor networks: evolution, opportunities, and challenges. *Proceedings of the IEEE*, 91(8):1247–1256, 2003. doi:10.1109/JPROC.2003.814918.

[6] M. Younis and K. Akkaya. Strategies and techniques for node placement in wireless sensor networks: A survey. *Ad Hoc Networks*, 6(4):621–655, 2008. doi:10.1016/j.adhoc.2007.05.003.

[7] X. Z. Cheng, D. Z. Du, L. S. Wang, and B. G. Xu. Relay sensor placement in wireless sensor networks. *Wireless Networks*, 14(3):347–355, 2008. doi:10.1007/s11276-006-0724-8.

[8] K. Akkaya and M. Younis. Cola: A coverage and latency aware actor placement for wireless sensor and actor networks. In *IEEE Vehicular Technology Conference*, pages 1–5. doi:10.1109/VTCF.2006.544.

[9] W. Youssef and M. Younis. Intelligent gateways placement for reduced data latency in wireless sensor networks. In *2007 IEEE International Conference on Communications*, pages 3805–3810. doi:10.1109/ICC.2007.627.

[10] F. Susanto, S. Budi, P. de Souza, U. Engelke, and J. He. Design of environmental sensor networks using evolutionary algorithms. *IEEE Geoscience and Remote Sensing Letters*, 13(4):575–579, 2016. doi:10.1109/LGRS.2016.2525980.

[11] M. Mansouri, H. Nounou, and M. Nounou. Genetic algorithm-based adaptive optimization for target tracking in wireless sensor networks. *Journal of Signal Processing Systems*, 74(2):189–202, 2013. doi:10.1007/s11265-013-0758-y.

[12] S. Budi, P. de Souza, G. Timms, V. Malhotra, and P. Turner. Optimisation in the design of environmental sensor networks with robustness consideration. *Sensors*, 15(12):29765–81, 2015. doi:10.3390/s151229765.

[13] J. H Wang, G. Yong, G. B. M. Heuvelink, and C. H. Zhou. Spatial sampling design for estimating regional gpp with spatial heterogeneities. *Geoscience and Remote Sensing Letters, IEEE*, 11(2):539–543, 2014. doi:10.1109/LGRS.2013.2274453.

[14] G. B. M. Heuvelink, Z. Jiang, S. De Bruin, and C. J. W. Twenhöfel. Optimization of mobile radioactivity monitoring networks. *International Journal of Geographical Information Science*, 24(3):365–382, 2010. doi:10.1080/13658810802646687.

[15] Y. Ge, J. H. Wang, G. B. M. Heuvelink, R. Jin, X. Li, and J. F. Wang. Sampling design optimization of a wireless sensor network for monitoring ecohydrological processes in the babao river basin, china. *International Journal of Geographical Information Science*, 29(1):92–110, 2015. doi:10.1080/13658816.2014.948446.

[16] J. Kang, X. Li, R. Jin, Y. Ge, J. Wang, and J. Wang. Hybrid optimal design of the eco-hydrological wireless sensor network in the middle reach of the heihe river basin, china. *Sensors (Basel)*, 14(10):19095–114, 2014. doi:10.3390/s141019095.

[17] A. Frery, H. S. Ramos, J. Alencar-Neto, E. Nakamura, and A. A. F. Loureiro. Data driven performance evaluation of wireless sensor networks. *Sensors*, 10(3):2150, 2010.

[18] P. Cheng, C. N. Chuah, and X. Liu. Energy-aware node placement in wireless sensor networks. In *IEEE Global Telecommunications Conference, 2004.*, volume 5, pages 3210–3214 Vol.5. doi:10.1109/GLOCOM.2004.1378943.

[19] K. Xu, Q. Wang, H. Hassanein, and G. Takahara. Optimal wireless sensor networks (wsns) deployment: minimum cost with lifetime constraint. In *IEEE International Conference on Wireless And Mobile Computing, Networking And Communications, 2005.*, volume 3, pages 454–461 Vol. 3. doi:10.1109/WIMOB.2005.1512937.

[20] J. Tang, B. Hao, and A. Sen. Relay node placement in large scale wireless sensor networks. *Computer Communications*, 29(4):490–501, 2006. doi:10.1016/j.comcom.2004.12.032.

[21] T. Clouqueur, V. Phipatanasuphorn, P. Ramanathan, and K. K. Saluja. Sensor deployment strategy for target detection. In *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*, pages 42–48, 570745, 2002. ACM. doi:10.1145/570738.570745.

[22] D. Ganesan, R. Cristescu, and B. Beferull-Lozano. Power-efficient sensor placement and transmission structure for data gathering under distortion constraints. *ACM Trans. Sen. Netw.*, 2(2):155–181, 2006. doi:10.1145/1149283.1149284.

[23] N. Heo and P. K. Varshney. Energy-efficient deployment of intelligent mobile sensor networks. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 35(1):78–92, 2005. doi:10.1109/TSMCA.2004.838486.

[24] G. Wang, G. H. Cao, and T. La Porta. Movement-assisted sensor deployment. In *Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, volume 4, pages 2469–2479 vol.4. doi:10.1109/INFCOM.2004.1354668.

[25] J. Wu and S. Yang. Smart: a scan-based movement-assisted sensor deployment method in wireless sensor networks. In *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, volume 4, pages 2313–2324 vol. 4. doi:10.1109/INFCOM.2005.1498518.

[26] M. Younis, M. Bangad, and K. Akkaya. Base-station repositioning for optimized performance of sensor networks. In *Vehicular Technology Conference, 2003-Fall. IEEE 58th*, volume 5, pages 2956–2960 Vol.5. doi:10.1109/VETECF.2003.1286165.

[27] I. Vasilescu, K. Kotay, D. Rus, M. Dunbabin, and P. Corke. Data collection, storage, and retrieval with an underwater sensor network, 2005. doi:10.1145/1098918.1098936.

[28] P. Sikka, P. Corke, and L. Overs. Wireless sensor devices for animal tracking and control, 2004. URL: http://eprints.qut.edu.au/33837/, doi:10.1109/lcn.2004.141.

[29] H. F. Abou-Shaara. The foraging behaviour of honey bees, *Apis mellifera*: a review. *Vet Med (Praha)*, 59(1):1–10, 2014.

[30] E. A. Capaldi, A. D. Smith, J. L. Osborne, S. E. Fahrbach, S. M. Farris, D. R. Reynolds, A. S. Edwards, A. Martin, G. E. Robinson, G. M. Poppy, and J. R. Riley. Ontogeny of orientation flight in the honeybee revealed by harmonic radar. *Nature*, 403(6769):537–540, 2000.

[31] H. de Vries and J. C. Biesmeijer. Modelling collective foraging by means of individual behaviour rules in honey-bees. *Behavioral Ecology and Sociobiology*, 44(2):109–124, 1998. doi:10.1007/s002650050522.

[32] J. R. Riley, U. Greggers, A. D. Smith, D. R. Reynolds, and R. Menzel. The flight paths of honeybees recruited by the waggle dance. *Nature*, 435(7039):205–207, 2005.

[33] A. M. Reynolds, A. D. Smith, D. R. Reynolds, N. L. Carreck, and J. L. Osborne. Honeybees perform optimal scale-free searching flights when attempting to locate a food source. *Journal of Experimental Biology*, 210(21):3763–3770, 2007. doi:10.1242/jeb.009563.

[34] J. L. Osborne, A. Smith, S. J. Clark, D. R. Reynolds, M. C. Barron, K. S. Lim, and A. M. Reynolds. The ontogeny of bumblebee flight trajectories: From naïve explorers to experienced foragers. *PLOS ONE*, 8(11):e78681, 2013. doi:10.1371/journal.pone.0078681.

[35] T. D. Seeley, S. Camazine, and J. Sneyd. Collective decision-making in honey bees: how colonies choose among nectar sources. *Behavioral Ecology and Sociobiology*, 28(4):277–290, 1991. doi:10.1007/bf00175101.

[36] B. Granovskiy, T. Latty, M. Duncan, D. J. T. Sumpter, and M. Beekman. How dancing honey bees keep track of changes: the role of inspector bees. *Behavioral Ecology*, 23(3):588–596, 2012. doi:10.1093/beheco/ars002.

[37] G. E. Robinson, B. A. Underwood, and C. E. Henderson. A highly specialized water-collecting honey bee. *Apidologie*, 15(3):355–358, 1984.

[38] C. W. W. Pirk, J. R. de Miranda, M. Kramer, T. E. Murray, F. Nazzi, D. Shutler, J. J. M. van der Steen, and C. van Dooremalen. Statistical guidelines for *Apis Mellifera* research. *Journal of Apicultural Research*, 52(4):1–24, 2013. doi:10.3896/IBRA.1.52.4.13.

[39] A. Van Geystelen, K. Benaets, D. C. de Graaf, M. H. D. Larmuseau, and T. Wenseleers. Track-a-forager: a program for the automated analysis of rfid tracking data to reconstruct foraging behaviour. *Insectes Sociaux*, 63(1):175–183, 2016. doi:10.1007/s00040-015-0453-z.

[40] J. J. Garcia Adeva. Simulation modelling of nectar and pollen foraging by honeybees. *Biosystems Engineering*, 112(4):304–318, 2012. doi:10.1016/j.biosystemseng.2012.05.002.

[41] J. Li and A. D. Heap. Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling and Software*, 53(0):173–189, 2014. doi:10.1016/j.envsoft.2013.12.008.

[42] P. A Burrough and R. A McDonnell. *Principles of geographical information Systems*. Oxford University Press, Oxford, 1998.

[43] C. V. Deutsch. Correcting for negative weights in ordinary kriging. *Computers and Geosciences*, 22(7):765–773, 1996. doi:10.1016/0098-3004(96)00005-2.

[44] L. Li and P. Revesz. *A Comparison of Spatio-temporal Interpolation Methods*, pages 145–160. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002. ISBN: 978-3-540-45799-2. doi:10.1007/3-540-45799-2_11.

[45] L. Li. Spatiotemporal interpolation methods in GIS, 2003. URL: http://digitalcommons.unl.edu/dissertations/AAI3092570.

[46] D. Shepard. A two-dimensional interpolation function for irregularly-spaced data, 1968. doi:10.1145/800186.810616.

[47] D. S. Shepard. *Computer Mapping: The SYMAP Interpolation Algorithm*, volume 40 of *Theory and Decision Library*, book section 7, pages 133–145. Springer Netherlands, 1984. doi:10.1007/978-94-017-3048-8_7.

[48] J. Li and A. D. Heap. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecological Informatics*, 6(3–4):228–241, 2011. doi:10.1016/j.ecoinf.2010.12.003.

[49] F. W. Chen and C. W. Liu. Estimation of the spatial rainfall distribution using inverse distance weighting (idw) in the middle of Taiwan. *Paddy and Water Environment*, 10(3):209–222, 2012. doi:10.1007/s10333-012-0319-1.

[50] L. de Mesnard. Pollution models and inverse distance weighting: Some critical remarks. *Computers and Geosciences*, 52:459–469, 2013. doi:10.1016/j.cageo.2012.11.002.

[51] L. Li, T. Losser, C. Yorke, and R. Piltner. Fast inverse distance weighting-based spatiotemporal interpolation: A web-based application of interpolating daily fine particulate matter pm2.5 in the contiguous u.s. using parallel programming and k-d tree. *International Journal of Environmental Research and Public Health*, 11(9):9101–9141, 2014. doi:10.3390/ijerph110909101.

[52] J. J. J. C. Snepvangers, G. B. M. Heuvelink, and J. A. Huisman. Soil water content interpolation using spatio-temporal kriging with external drift. *Geoderma*, 112(3–4):253–271, 2003. doi:10.1016/S0016-7061(02)00310-5.

[53] S. De Iaco, D. E. Myers, and D. Posa. Space–time analysis using a general product–sum model. *Statistics & Probability Letters*, 52(1):21–28, 2001. doi:10.1016/S0167-7152(00)00200-5.

[54] J. Li. *A review of spatial interpolation methods for environmental scientists / Jin Li and Andrew D. Heap*. Record (Australia. Geoscience Australia) ; 2008/23. Geoscience Australia, Canberra, 2008. URL: http://www.ga.gov.au/servlet/BigObjFileManager?bigobjid=GA12526.

[55] J. R. Shewchuk. Tetrahedral mesh generation by delaunay refinement, 1998. doi:10.1145/276884.276894.

[56] L. A. Freitag and C. Ollivier-Gooch. Tetrahedral mesh improvement using swapping and smoothing. *International Journal for Numerical Methods in Engineering*, 40(21):3979–4002, 1997. doi:10.1002/(SICI)1097-0207(19971115)40:21<3979::AID-NME251>3.0.CO;2-9.

[57] S. M. Pingale, D. Khare, M. K. Jat, and J. Adamowski. Spatial and temporal trends of mean and extreme rainfall and temperature for the 33 urban centers of the arid and semi-arid state of rajasthan, india. *Atmospheric Research*, 138(0):73–90, 2014. doi:10.1016/j.atmosres.2013.10.024.

[58] G. Y. Lu and D. W. Wong. An adaptive inverse-distance weighting spatial interpolation technique. *Computers and Geosciences*, 34(9):1044–1055, 2008. doi:10.1016/j.cageo.2007.07.010.

[59] V. Roshan Joseph and Lulu Kang. Regression-based inverse distance weighting with applications to computer experiments. *Technometrics*, 53(3):254–265, 2011. doi:10.1198/TECH.2011.09154.

[60] M. New, M. Hulme, and P. Jones. Representing twentieth-century space–time climate variability. part ii: Development of 1901–96 monthly grids of terrestrial surface climate. *Journal of Climate*, 13(13):2217–2238, 2000. doi:10.1175/1520-0442(2000)013<2217:RTCSTC>2.0.CO;2.

[61] J. Caesar, L. Alexander, and R. Vose. Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set. *Journal of Geophysical Research: Atmospheres*, 111(D5):D05101, 2006. doi:10.1029/2005JD006280.

[62] D. Kiktev, D. M. H. Sexton, L. Alexander, and C. K. Folland. Comparison of modeled and observed trends in indices of daily climate extremes. *Journal of Climate*, 16(22):3560–3571, 2003. doi:10.1175/1520-0442(2003)016<3560:COMAOT>2.0.CO;2.

[63] S. J. Jeffrey, J. O. Carter, K. B. Moodie, and A. R. Beswick. Using spatial interpolation to construct a comprehensive archive of australian climate data. *Environmental Modelling and Software*, 16(4):309–330, 2001. doi:10.1016/S1364-8152(01)00008-1.

[64] M. R. Haylock, N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, and M. New. A european daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *Journal of Geophysical Research: Atmospheres*, 113(D20):D20119, 2008. doi:10.1029/2008JD010201.

[65] S. Naoum and I. K. Tsanis. Ranking spatial interpolation techniques using a gis-based dss. *Global Nest*, 6(1):1–20, 2004. doi:10.1016/j.egypro.2012.05.058.

[66] C. C. F. Plouffe, C. Robertson, and L. Chandrapala. Comparing interpolation techniques for monthly rainfall mapping using multiple evaluation criteria and auxiliary data sources: A case study of sri lanka. *Environmental Modelling and Software*, 67(0):57–71, 2015. doi:10.1016/j.envsoft.2015.01.011.

[67] C. A. Gotway, R. B. Ferguson, G. W. Hergert, and T. A. Peterson. Comparison of kriging and inverse-distance methods for mapping soil parameters. *Soil Science Society of America Journal*, 60(4):1237–1247, 1996. doi:10.2136/sssaj1996.03615995006000040040x.

[68] M. Meul and M. Van Meirvenne. Kriging soil texture under different types of nonstationarity. *Geoderma*, 112(3–4):217–233, 2003. doi:10.1016/S0016-7061(02)00308-7.

[69] C. A. Schloeder, N. E. Zimmerman, and M. J. Jacobs. Comparison of methods for interpolating soil properties using limited data. *Soil Science Society of America Journal*, 65(2):470–479, 2001. URL: 10.2136/sssaj2001.652470x, doi:10.2136/sssaj2001.652470x.

[70] Y. Sun, S. Kang, F. Li, and L. Zhang. Comparison of interpolation methods for depth to groundwater and its temporal and spatial variations in the minqin oasis of northwest china. *Environmental Modelling and Software*, 24(10):1163–1170, 2009. doi:10.1016/j.envsoft.2009.03.009.

[71] J. P. Matos, T. Cohen Liechti, D. Juízo, M. M. Portela, and A. J. Schleiss. Can satellite based pattern-oriented memory improve the interpolation of sparse historical rainfall records? *Journal of Hydrology*, 492:102–116, 2013. doi:10.1016/j.jhydrol.2014.01.003.

[72] J. P. Matos, T. Cohen Liechti, M. M. Portela, and A. J. Schleiss. Pattern-oriented memory interpolation of sparse historical rainfall records. *Journal of Hydrology*, 510:493–503, 2014. doi:10.1016/j.jhydrol.2014.01.003.

[73] L. A. Sanabria, X. Qin, J. Li, R. P. Cechet, and C. Lucas. Spatial interpolation of mcarthur's forest fire danger index across australia: Observational study. *Environmental Modelling and Software*, 50(0):37–50, 2013. doi:10.1016/j.envsoft.2013.08.012.

[74] L. Li and P. Revesz. Interpolation methods for spatio-temporal geographic data. *Computers, Environment and Urban Systems*, 28(3):201–227, 2004. doi:10.1016/S0198-9715(03)00018-8.

[75] L. Li, X. Zhang, J. B. Holt, J. Tian, and R. Piltner. Spatiotemporal interpolation methods for air pollution exposure. In *Proceedings of the Ninth Symposium on Abstraction, Reformulation and Approximation*. AAAI. URL: https://www.aaai.org/ocs/index.php/SARA/SARA11/paper/view/4241.

[76] M. Kilibarda, T. Hengl, G. B. M. Heuvelink, B. Graler, E. Pebesma, M. Perčec Tadić, and B. Bajat. Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. *Journal of Geophysical Research: Atmospheres*, 119(5):2294–2313, 2014. doi:10.1002/2013JD020803.

[77] Z. Zeng, L. Lei, S. Hou, and L. Li. A spatio-temporal interpolation approach for the FTS SWIR product of XCO2 data from GOSAT. In *Geoscience and Remote Sensing Symposium, 2012 IEEE International*, pages 852–855. doi:10.1109/IGARSS.2012.6351427.

[78] M. A. Becher, V. Grimm, P. Thorbek, J. Horn, P. J. Kennedy, and J. L. Osborne. Beehave: a systems model of honeybee colony dynamics and foraging to explore multifactorial causes of colony failure. *Journal of Applied Ecology*, 51(2):470–482, 2014. doi:10.1111/1365-2664.12222.

[79] D. Zimmerman, C. Pavlik, A. Ruggles, and M. P. Armstrong. An experimental comparison of ordinary and universal kriging and inverse distance weighting. *Mathematical Geology*, 31(4):375–390, 1999. URL: 10.1023/A:1007586507433, doi:10.1023/A:1007586507433.

[80] P. de Souza, R. Williams, S. Quarrell, S. Budi, F. Susanto, B. Vincent, G. Allen, A. Almeida, D. Worledge, L. Disiuta, P. Hirsch, G. Pessin, H. Arruda, P. Marendy, L. dos Santos, T. Gillard, and A. O. Ong. Agent-based modelling of honey bee forager flight behaviour for swarm sensing applications, 2016. (under review).

[81] J. Katzfey and M. Thatcher. Ensemble one-kilometre forecasts for the south esk hydrological sensor web, 12-16 December 2011. URL: http://www.mssanz.org.au.previewdns.com/modsim2011/I5/katzfey.pdf.

[82] R. K. Rew, G. P. Davis, and S. Emmerson. Netcdf user's guide, an interface for data access, version 2.3, 1993. URL: http://www.unidata.ucar.edu/software/netcdf/docs/user_guide.html.

[83] U. Engelke, F. Susanto, P. A. de Souza, and P. Marendy. Visual assessment of spatial data interpolation. In *Big Data Visual Analytics, 2015*, pages 1–7. doi:10.1109/BDVA.2015.7314305.

[84] L. Tang, X. Su, G. Shao, H. Zhang, and J. Zhao. A clustering-assisted regression (car) approach for developing spatial climate data sets in china. *Environmental Modelling and Software*, 38(0):122–128, 2012. doi:10.1016/j.envsoft.2012.05.008.

[85] J. Breen, P. de Souza, G. P. Timms, and R. Ollington. Onboard assessment of xrf spectra using genetic algorithms for decision making on an autonomous underwater vehicle. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 269(12):1341–1345, 2011. URL: http://www.sciencedirect.com/science/article/pii/S0168583X11003156, doi:10.1016/j.nimb.2011.03.012.

[86] F. Susanto, P. de Souza, and J. He. Spatiotemporal interpolation for environmental modelling. *Sensors (Basel)*, 16(8), 2016. doi:10.3390/s16081245.

[87] S. Anderson. An evaluation of spatial interpolation methods on air temperature in phoenix, az, 2002. Access Date: 13 May 2014. URL: http://www.cobblestoneconcepts.com/ucgis2summer/anderson/anderson.htm.

[88] S. van der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, 13(2):22–30, 2011. doi:10.1109/MCSE.2011.37.

[89] E. Jones, T. Oliphant, and P. Peterson. Scipy: Open source scientific tools for python, 2001–. [Online; accessed 2016-05-26]. URL: http://www.scipy.org.

[90] W. McKinney. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, pages 51 – 56.

[91] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi:10.1109/MCSE.2007.55.

[92] M. Waskom, O. Botvinnik, drewokane, P. Hobson, David, Y. Halchenko, S. Lukauskas, J. B. Cole, J. Warmenhoven, J. de Ruiter, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, M. Martin, A. Miles, K. Meyer, T. Augspurger, T. Yarkoni, P. Bachant, M. Williams, C. Evans, C. Fitzgerald, Brian, D. Wehner, G. Hitz, E. Ziegler, A. Qalieh, and A. Lee. seaborn: v0.7.1 (june 2016). 2016. doi:10.5281/zenodo.54844.

[93] F. A. Fortin, F. M. De Rainville, M. A. Gardner Gardner, M. Parizeau, and C. Gagné. Deap: evolutionary algorithms made easy. *J. Mach. Learn. Res.*, 13(1):2171–2175, 2012.